

ABSTRACT

Title of dissertation: REVEALING PERCEPTUAL PROXIES IN
COMPARATIVE DATA VISUALIZATION
Brian Ondov, Doctor of Philosophy, 2021

Dissertation directed by: Professor Niklas Elmqvist
College of Information Studies

Data Visualization has long been shaped by empirical evidence of the efficacies of different encodings, such as length, position, or area, in conveying quantities. Less is known, however, about what may affect comparison of multiple data series, which generally involves extraction of higher-order values, such as means, ranges, and correlations. In this work, we investigate such factors and the underlying visual processes that may account for them. We begin with a case study motivating the research, in which we modify Krona, a Bioinformatics visualization system, to support several types of comparison. Next, we empirically examine the influence of “arrangement”—that is, whether charts are shown side-by-side, stacked vertically, overlaid, etc.—on comparative tasks, in a series of psychophysical experiments. The results suggest a complex interaction of factors, with different comparative arrangements providing benefits for different combinations of tasks and encodings. For example, overlaid charts make detecting differences easier but comparing means or ranges more difficult. While these results offer some guidance to designers, the number of interactions makes it infeasible to provide broad rankings of arrangements, as

has been done previously for encodings. Our subsequent efforts thus work toward understanding the visual processes that underlie the extraction of statistical summaries needed for comparison. It has recently been proposed that simpler shortcuts, called Perceptual Proxies, are used by the visual system to estimate these values. We investigate proxies for bar charts in experiments using an “adversarial” framework, in which the ranking of two charts along a task metric (e.g. mean) is opposite their ranking along a proxy metric (e.g. convex hull area). The strongest evidence we find is for use of a “centroid” proxy to estimate means in bar charts. Finally, we attempt to use human-guided optimization to construct charts *de novo*, without assuming specific proxies. This work contributes both to perceptual psychology, by offering evidence for underlying visual processes that may be involved in the interpretation of comparative visualizations, and to data visualization, by providing new research methods and straightforward design guidance on how best to lay out charts to support certain tasks.

REVEALING PERCEPTUAL PROXIES IN COMPARATIVE DATA VISUALIZATION

by

Brian Ondov

Dissertation proposal submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2021

Advisory Committee:

Professor Niklas Elmqvist, Chair/Advisor

Professor Leilani Battle

Professor Eun Kyoung Choe

Professor John Dickerson

Professor Zhicheng Liu

Professor Rob Patro

Dr. Adam Phillippy

© Copyright by
Brian Ondov
2021

Acknowledgments

First, I would like thank my advisor, Professor Niklas Elmqvist, who had faith enough to take me on as a student after a single, brief meeting, and who, in an early brainstorming session, handed me an idea that would become the thread of this dissertation. Professor Elmqvist has been an ally and an advocate through the hurdles I've faced, and it is by his wisdom that I was able to wrangle my capricious interests enough to focus on completing this work.

Next, I would like to thank all the professors I've had the pleasure of learning from in the classroom. I recall, at the start of my degree, how crestfallen I was that only one course from my previous graduate studies could be transferred, and how much I dreaded the work that lay ahead of me. By the end, however, I didn't even need the transferred credits, engrossed as I was in the offerings of the faculty. Each of those professors pushed me to better myself, in a direct sense by upholding academic rigor, but also in a more transcendent sense, by embodying who and what I could become if I rose to the challenge.

Finally, I would like to thank my long-time supervisor, mentor, and friend, Adam Phillippy. I may never have pursued, let alone been accepted into, this doctoral program if it hadn't been for a partnership that has produced impact and reputation far beyond what I could have precipitated on my own. I am grateful for the confidence with which Doctor Phillippy gave me opportunities and for the patience with which he let me defect toward my own ideas. Whatever I may accomplish in the future will be owed in large part to this fortuitous relationship.

Table of Contents

Acknowledgements	ii
Table of Contents	iii
List of Figures	vi
1 Introduction	1
1.1 Contributions	7
1.2 Publications	8
1.3 Availability of Data and Implementations	9
2 Background	10
2.1 Visual Comparison	10
2.2 Perceptual Factors in Comparison	11
2.2.1 Co-location	12
2.2.2 Symmetry	12
2.2.3 Movement	13
2.3 Perceptual Proxies	14
2.4 Adversarial Visualization and Data	16
3 Case Study: Microbiome Comparison	18
3.1 Method	19
3.2 Task	19
3.3 Results	20
4 User Study Framework	23
4.1 Arrangements	24
4.2 Timed Impressions	26
4.3 Staircase Titration	27
4.4 Dynamic Data Generation	28
4.5 Rendering	29
4.6 Crowdsourcing	30
4.6.1 Training	30
4.6.2 Participant Recruitment and Payment	30

4.7	Procedure	31
5	Experiment 1: Maximum Delta Task	33
5.1	Experimental Setup	33
5.2	Data Generation	36
5.3	Hypotheses	37
5.4	Results	39
5.4.1	Exp. 1A: Bar charts	39
5.4.2	Exp. 1A Floor Effect	41
5.4.3	Exp. 1B: Slope charts	42
5.4.4	Exp. 1C: Donut charts	43
5.5	Discussion	44
6	Experiment 2: Correlation task	46
6.1	Experimental Setup	46
6.2	Data Generation	48
6.3	Results	48
6.4	Summary	49
7	Experiment 3: Maximum Mean Task	50
7.1	Experimental Setup	50
7.2	Data Generation	51
7.3	Results	53
7.4	Summary	55
8	Experiment 4: Maximum Range Task	56
8.1	Experimental Setup	56
8.2	Data Generation	57
8.3	Results	57
8.4	Summary	60
9	Perceptual Proxies	62
9.1	Candidate Proxies	63
9.1.1	Global Features	63
9.1.2	Focal Features	65
9.2	Testing Proxies with Retrospective Analysis	66
9.2.1	Implementation	67
9.2.2	Results	70
9.2.3	Limitations	74
10	Revealing Proxies with Adversarial Examples	75
10.1	Two Approaches: Testing vs. Learning	75
10.2	Common Methods For Adversarial Experiments	77
10.2.1	Visual Representation	77
10.2.2	Tasks	78
10.2.3	Procedure	79

10.2.4	Participants	80
10.2.5	Apparatus	80
11	Experiment 5: Testing Proxies with Adversarial Charts	82
11.1	Selecting Specific Proxies	82
11.2	Eliminating Confounding Proxies	84
11.3	Hypotheses	86
11.4	Experimental Design	86
11.5	Generating Adversarial Charts with Simulated Annealing	87
11.6	Measurement	88
11.7	Prerequisites for Analysis	90
11.8	Analysis	92
11.8.1	Step 1: Deriving Thresholds and Measurement Error	92
11.8.2	Step 2: Modeling Thresholds	94
11.9	Results	95
11.9.1	The Effects of Manipulating Perceptual Proxies	95
11.9.2	Interpreting Participants Selecting Against a Proxy	96
11.9.3	Individual Differences	98
12	Experiment 6: Learning Adversarial Charts Interactively	102
12.1	Optimization Method	103
12.2	Experimental Design	107
12.3	Analysis	108
12.4	Results	109
12.5	Discussion	111
13	Discussion	112
13.1	Implications for Data Visualization Practice	112
13.1.1	Design Guidance	112
13.1.2	Adversarial Visualizations and Deception	114
13.2	Implications for Data Visualization Research	115
13.3	Implications for Perceptual Psychology	116
13.4	Limitations	119
14	Future Work	121
14.1	Continuing to Solve the Cube	121
14.2	Generating New Candidate Proxies	122
14.3	Proxies Cubed	122
14.4	How Might Viewers Choose Proxies?	123
14.5	Automated Systems	123
15	Conclusion	125
	Bibliography	127

List of Figures

1.1	Arrangement as a third dimension of evaluation	2
1.2	Examples of comparative evaluations	4
3.1	Comparative modes implemented in Krona	22
4.1	Comparative arrangement methods examined	24
4.2	Examples of the mirrored arrangement	25
4.3	The staircase titration method	27
4.4	The ready screen for trials	31
5.1	Encodings used for the Maximum Delta task	34
5.2	Response prompt for Maximum Delta task	35
5.3	Results for the Maximum Delta task with bar charts	39
5.4	Results for the Maximum Delta task with slope charts	40
5.5	Results for the Maximum Delta task with donut charts	41
5.6	Titer histograms for Maximum Delta (bar charts)	43
6.1	Example renderings of the Correlation task	47
6.2	Results for the Correlation task	48
7.1	Response prompt for Maximum Mean task	51
7.2	Results for the Maximum Mean task	54
8.1	Results for the Maximum Range task	59
9.1	Candidate global perceptual proxies	64
9.2	Candidate focal perceptual proxies	65
9.3	Geometric encoding of a bar chart	68
9.4	Comparison of proxies to human choices (MaxMean)	71
9.5	Comparison of proxies to human choices (MaxRange)	72
9.6	Proxy correlation for Maximum Mean data	73
9.7	Proxy correlation for Maximum Range data	74
10.1	Conceptual diagram of two adversarial approaches	76
10.2	Interleaved experimental procedure	78

11.1	The confounding proxies in the <i>MaxMean</i> and <i>MaxRange</i> tasks.	83
11.2	Perceptual proxies used for <i>MaxMean</i> adversarial experiments	84
11.3	Perceptual proxies used for <i>MaxRange</i> adversarial experiments	85
11.4	Example titers for <i>MaxMean</i>	89
11.5	Example titers for <i>MaxRange</i>	90
11.6	Deriving titer thresholds and measurement error	93
11.7	The effects of manipulating perceptual proxies for <i>MaxMean</i>	96
11.8	The effects of manipulating perceptual proxies for <i>MaxRange</i>	97
11.9	An example of participants selecting against a proxy.	98
11.10	Individual differences in adversarial mean trials	100
11.11	Individual differences in adversarial range trials	101
12.1	Quantitative analyses of learned charts for <i>MaxMean</i>	109
12.2	Quantitative analyses of learned charts for range	110

Chapter 1: Introduction

While the visualization designer has myriad ways to represent information graphically, experimental evaluation has shown us that not all representations are equal [1–3]. These perceptual studies are often motivated by tasks that are typical for analyzing a single data series, e.g. averages, trends, extreme values, and outliers [4]. When comparing more than one dataset, however, the goals of the visualization can be fundamentally different [5]. For example, instead of looking for the largest or smallest data point, we may look for the largest *delta* from one set to another [6], or for an overall level of correlation [7]. Further, we may need to extract a summary statistic, such as the mean or range from each chart in a group of charts in order to compare them. While many of the perceptual lessons learned from single series no doubt extend to these tasks, introducing comparison can tax substantially capacity-limited aspects of our visual system, such as abstract object representations and the selection of those representations [8]. We thus posit that, in addition to studying the influence of encoding (e.g. Cleveland & McGill [2] and Simkin & Hastie [9] and task (e.g. Kim & Heer [10] and Amar & Stasko [11]), it will also be valuable to consider *arrangement* as a third dimension of the factors that specifically affect the efficacy of comparative visualization, as depicted in Figure 1.1.

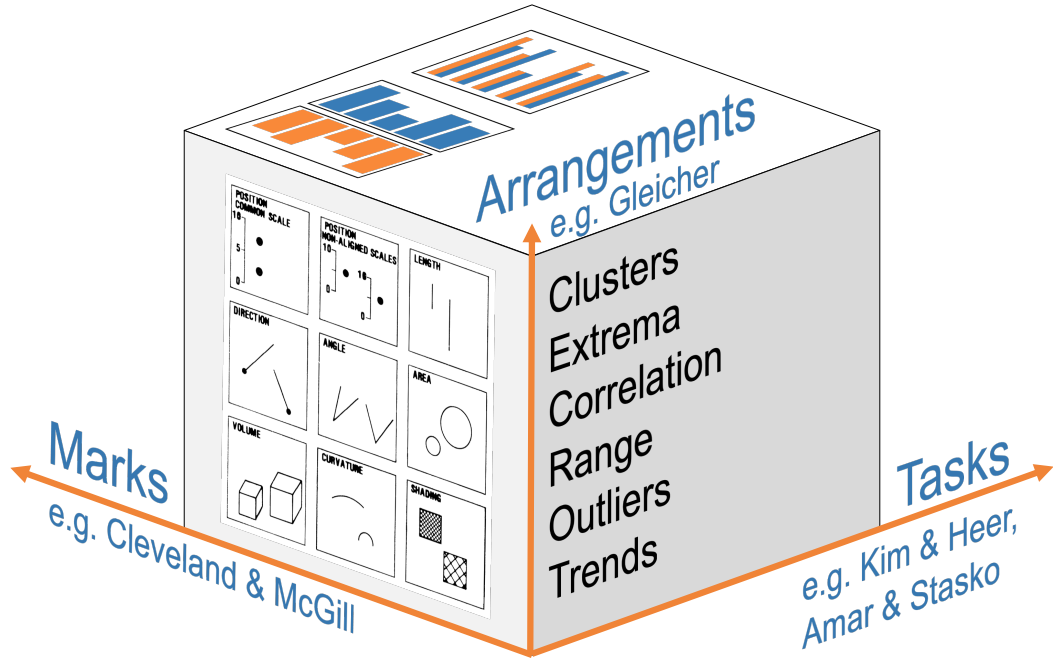


Figure 1.1: Visual comparison depends not on a single dimension of mark, arrangement, or task, but of the interactions between them. These interactions can be represented as a cube. Our present goal is not to examine the full space of the cube, but rather to understand how a viewer uses visual features to serve analytic task goals depending on the marks and arrangements they see.

In the vein of prior work on elementary encodings, this work will seek to elucidate what makes comparative displays effective and to offer guidance for maximizing their efficacy. We begin with a case study of visual comparison within a taxonomic hierarchy browser, called Krona [12], based on sunburst displays [13] (Figure 1.2, right). We then present a series of graphical perception experiments designed to evaluate designs for visual comparison tasks.

We choose four primitive tasks specific to the goals of comparison: (1) identification of a maximum delta (or “biggest mover”) between data series, (2) estimation of overall correlation between two series, (3) comparison of mean values of two data series, and (4) comparison of ranges of two data series. These tasks are motivated by the low-level analytic task taxonomy of Amar et al. [7] and intended to be diverse in terms of their compositional modalities. For example, Task 1 (maximum delta, or “biggest mover”) requires a series of pairwise difference estimates across the charts followed by the extraction of a maximum from the resulting values (or potentially detection of an outlier, depending on the distribution of those values). Task 2 (correlation), however, is a single, primitive task in the taxonomy, and examines a pair of charts holistically. Tasks 3 and 4 (mean and range, respectively) both require the extraction of a single, summary value for each chart, which are then compared. While seeking these summary values may seem contrived in themselves, both are described by Amar et al. as building blocks for deeper tasks. For example, they cite the mean being used to compare relative efficiencies of two categories of cars, or ranges being used to assess whether a data series could merit further analysis.

We embed Task 1 in various stimuli (Figure 1.2, center): (a) length, represented as bar charts, (b) slope, represented as simple line graphs, and (c) angle, represented as donut charts. We embed Task 2 in a forced-choice between two *pairs* of bar charts (Figure 1.2, left). We embed Task 3 and 4 in a forced-choice between two *individual* bar charts. For each embedding, we explore performance of 5 arrangements: (i) ‘stacked’ small multiples with a common baseline, (ii) ‘adjacent’

small multiples with a non-common baseline [14],¹ (iii) superposition, or ‘overlaid’ charts, (iv) adjacent small multiples that are mirror symmetric, and (v) animated transitions. The first three of these are commonly used and are associated with intuitive—but rarely measured—differences in efficacy [17]. The last two are less common but may leverage the visual system’s sensitivity to motion [18], and in particular common motion [19], in addition to the sensitivity of the visual system to mirror symmetry of objects [20], making them valuable to evaluate.

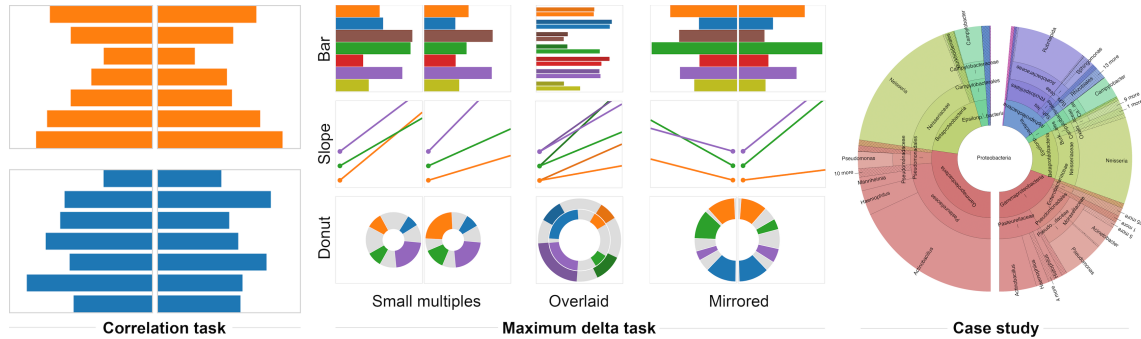


Figure 1.2: Evaluation methods for visual comparison. **Left:** Participants were asked to pick the most similar pair of bar charts for a variety of arrangements and degrees of correlation. **Center:** Participants were asked to find the maximum delta, or “biggest mover,” between pairs of datasets. Additional arrangements not shown are vertical small multiples and animated transitions. **Right:** Domain experts were interviewed after trying various comparative arrangements in Krona, an interactive sunburst display for biological data.

We find that ability to perform the tasks, as measured by the difficulty level need to achieve 75% accuracy in forced-choice experiments, is not optimized by a

¹We only examine a subset of (i) and (ii) for donut charts, as they have no inherent orientation. However, recent work on performance asymmetries between vertical vs. horizontal display layouts [15, 16] suggests that this case merits future study.

single mark type or spatial arrangement. Instead, the precision of visual comparison depends on an interaction of mark type, arrangement, and task (Figure 1.1). The best static chart for a precise delta comparison, for example, was one that was spatially superposed (“overlaid”), rather than juxtaposed, validating an intuitively-motivated guideline from Gleicher et al. [21]. Surprisingly, however, in some cases we also find significant task performance improvements when arranging small multiples in a mirror-symmetric fashion. Furthermore, counter to many prior studies showing animation to be ineffective in encoding quantitative information [6, 22, 23], we observe animation having high performance for the task of determining the datum with the biggest difference across two charts (“maximum delta” or “biggest mover”). Comparison of means and ranges, however, was most precise when the two datasets were vertically stacked, and least precise when the datasets were superposed. This pattern of which arrangements were best was strikingly different than for the previous pattern for tasks 1 and 2.

Why is there not a single clean emerging answer, where a given arrangement is best across various tasks? This empirical evidence for the more complex nature of visual comparison is consistent with the idea that it requires a series of visual actions at a variety of scales from one object (such as a single mark in a chart), to multiple objects, to whole sets of objects, such as entire charts in a small multiples setting [5]. Taxonomies of visual comparison describe multiple stages of perceptual and cognitive steps [4, 5], and vary in describing one or many types of visual comparisons, but the visual mechanisms supporting these stages are unclear. We argue that an empirical description of the precision of visual comparison across each com-

bination of mark \times arrangement \times task would be valuable, but unlikely to scale to have predictive value beyond its status as a lookup table. A different approach is required. Recently, the concept of *perceptual proxies*, which are theorized “shortcuts” that the visual system may take instead of computing statistics, has been gaining traction [24–26]. We propose that instead of continuing to fill out the entries of the cube in Figure 1.1, it may be more productive to study how perceptual proxies of a visualization are actually used to reason about a visual comparison task.

Drawing from perceptual psychology, as well as from data visualization and geometry, we compile a diverse, though by no means comprehensive, list of candidate perceptual proxies. Using trial data from our experiments investigating arrangement, we assess the plausibility of these proxies by comparing them to actual human choices. This lets us narrow down to a smaller set of proxies for further empirical study, with some representing broader classes of very similar proxies.

The fundamental problem for studying proxies empirically, however, is that, by definition, a plausible proxy should correlate well with the value a viewer is actually seeking. For example, if a viewer seeks the mean value of a series, a proxy with no relation to the mean is not one the viewer realistically could be using, as we know that people are fairly good at this task. How, then, can we ever know whether a viewer is using a particular proxy, rather than computing the true value, or using some other proxy?

Our solution to this apparent paradox is to try to use proxies to deceive participants when they are performing a task. This *adversarial* approach to testing proxies draws inspiration from the field of Machine Learning. It has long been

known that statistically learned models, such as deep artificial neural networks for computer vision, are subject to “adversarial attacks,” in which inputs can be manipulated to change classifier output despite looking very similar or identical to a human observer [27–30]. This phenomenon arises from the fact that these models, though somewhat analogous to our visual system, ultimately rely on very different representations of their input data. The human vision system, in turn, does not always process its input in the ways we might expect. This is evidenced by an abundance of optical illusions [31], which can be thought of as adversarial examples for that system. Just as carefully crafted illusions can be illustrative of underlying visual mechanisms, we hypothesize that manipulating data visualizations along particular axes can help reveal how they are interpreted. The perceptual proxies we have discussed will serve as those axes. Within this conceptual framework, we approach the problem experimentally in two complementary ways: (1) starting from proxies as assumptions and attempting to validate them, and (2) starting without assumptions and attempting to recover those proxies or discover new ones.

1.1 Contributions

The main contributions of this work are:

1. A case study in which we interview domain experts given various comparative modes implemented within the same platform (Krona [12]).
2. Results from four graphical perception experiments measuring participant performance across comparative arrangements for (a) a maximum delta task in

- bar, slope, and donut charts, (b) a correlation task for bar charts, (c) a maximum mean task for bar charts, and (d) a maximum range task for bar charts.
3. Data generation procedures designed specifically for graphical perception studies on visual comparison.
 4. A list of candidate perceptual proxies for visual comparison in bar charts and a cursory assessment of their plausibility using data from human subjects experiments on comparative arrangements.
 5. A framework for testing perceptual proxies using “adversarial examples” based on those proxies, and the results of experiments using this framework for maximum mean and maximum range tasks for bar charts.
 6. A framework for creating adversarial charts *de novo* using human-guided optimization, and the results of implementing this framework for maximum mean and maximum range tasks for bar charts.

1.2 Publications

This document includes work from the following peer-reviewed publications:

- Brian D. Ondov, Nicole Jardine, Niklas Elmqvist, and Steven Franconeri. Face to face: Evaluating visual comparison. *IEEE Transactions on Visualization and Computer Graphics*, 25(1):861–871, 2019
- Nicole Jardine, Brian D Ondov, Niklas Elmqvist, and Steven Franconeri. The perceptual proxies of visual comparison. *IEEE transactions on visualization*

and computer graphics, 26(1):1012–1021, 2019

- Brian D Ondov, Fumeng Yang, Matthew Kay, Niklas Elmqvist, and Steven Franconeri. Revealing perceptual proxies with adversarial examples. *IEEE Transactions on Visualization and Computer Graphics*, 2020

This work was highly collaborative and credit is due to all authors of these publications. BDO designed and implemented the experimental procedures and conducted all crowdsourced experiments. NE advised and wrote on the larger impact of the work on data visualization. SF and NJ advised and wrote on perceptual psychology and created the initial list of proxies. NJ performed analysis for Experiments 1-4, including written results and figures. MK advised on Bayesian analysis for Experiments 5 and 6. FY performed analysis for Experiments 5 and 6, including written results and figures.

1.3 Availability of Data and Implementations

In the interest of auditing and continued research we provide experimental implementations, data collected (anonymized as appropriate), and analysis scripts at <https://osf.io/yhxuz/> (Experiments 1 and 2), <https://osf.io/uenzd/> (Experiments 3 and 4), and <https://osf.io/2re7b/> (Experiments 5 and 6).

Chapter 2: Background

It is not enough to make visualizations that are pleasing or engaging—empirical evaluation is a crucial part of the analytical process [34]. Cleveland & McGill informed decades of design by ranking basic visual channels by their quantitative accuracy [2]. Specific visual faculties, like the detection of outliers and salient elements, have been also been well studied [35–38], and the widespread application of color theory to visualization has helped designers avoid skewed interpretations [39–41]. These types of studies typically involve relationships within a single data series, with tasks such as estimating size differences [42] or determining if points in the series are equal [43]. Often, however, real data are not so simple, requiring more complex comparisons across multiple series [21].

2.1 Visual Comparison

Expanding from a single data series to multiple constitutes a multivariate analysis, i.e. adding rows to a table in Bertin’s synoptic [44]. Comparative visualization (e.g. small multiples) can be thought of as multivariate analysis in which a *categorical* variable is used to slice the data. For example, we may want to compare time series of the popularity of various baby names or the prices of a variety of goods in

different countries. The goals of comparison are often different than those of single-series analysis and can be described as compounds of more primitive tasks [45]. Gleicher et al. provide taxonomies of tasks, as well as comprehensive reviews of techniques and best-practice guidance, specifically for comparative displays [5, 21]. While these reviews provide valuable intuition about the efficacy of various comparative strategies, quantitative user studies are less common in this area. Qu et al. explore the importance of consistent scale and coloring across small multiple displays, but not the efficacy of the arrangements themselves [46, 47]. Roberston et al. compare animation to a relatively high number of small multiples (8 to 80) for conveying trends in GapMinder data [6, 48]. Heer et al. compare variants of time-series representations within the context of vertical juxtaposition [49]. Javed et al. also evaluate various methods of displaying multiple time series and include both juxtaposition and superposition, but with tasks similar to those of single-view evaluations [50].

2.2 Perceptual Factors in Comparison

We consider here three themes from the perceptual psychology literature in considering which comparison arrangements to evaluate. This is not an exhaustive list of the factors that may be relevant, but will serve as a basis for experimentation.

2.2.1 Co-location

Within a single region of space, visual features such as length, orientation, and motion can rapidly convey information about stimulus deltas. Comparison between two regions is a more difficult task for the observer, because it may require an active process of storage of one region before being able to compare it with another region. “Spot the difference” games, in which observers try to detect small changes between two otherwise identical images, illustrate the difficulty of this task. Mental storage capacity, even for basic visual features like shapes and colors, is around four at maximum [51], and observer comparisons between mentally stored features and currently visible features may be subject to multiple bottlenecks [52]. Detecting a difference between two sets of data may only be possible for large change sizes, even for small datasets (e.g., 5-10 values).

2.2.2 Symmetry

An additional consideration for multiple displays is that the human visual system is sensitive to symmetry, and especially mirror symmetry located at the focal point [20, 53]. Specifically, the system’s ability to detect visual differences is more efficient between two regions that are otherwise mirror images of each other, compared to repeated translations of each other [54, 55] and when the symmetric information is spatially close rather than far [56]. Juxtaposed datasets (e.g. small multiples) are typically translated horizontally, and with common axial directions in order to reduce the cognitive burden of understanding the different polarities of

each side of the horizontal axes [5]. But mirror symmetry is occasionally used when comparing two data series that are similar, for example in population pyramids [57], suggesting that designers have an implicit awareness that this arrangement may hold benefits. We hypothesize that advantages for human symmetry detection could convey benefits for comparisons of data in mirrored arrangements.

2.2.3 Movement

Motion is a primitive and fundamental element of vision [18]. Estimates of velocity can originate in the retina itself [58], and at higher levels of visual processing motion can be used to extract statistics and structure from scenes [19, 59], and may be a useful cue for statistical extraction of patterns in data visualizations [60].

But motion processing is not all-powerful. In particular, when a viewer is asked to process multiple moving objects simultaneously, performance can fall drastically for more than 2-4 objects [61, 62]. When used to demonstrate processes in diagrams in teaching, its use can confuse students [63]. Evidence for the usefulness of motion in visualization is early and mixed. Animation can fill a wide variety of roles and may have similarly varied utility [64], and has shown promise in the role of maintaining context during configurational changes [65–68]. Because the visual system encodes motion speed and direction as a primitive and direct feature similar to orientation or length [18], it may be especially useful for detecting changes to values, because larger changes should co-vary with motion speed, and change direction with motion direction. Prior studies have assigned animation questionable value in similar tasks,

for example when conveying correlation via oscillation [22], conveying trends in time series [6], or linking two views in a scatterplot [23]. However, these are specific instances among a wide variety of possible tasks, graphical representations, and layouts.

2.3 Perceptual Proxies

One way to think about human vision is that it is an *information processing system* capable of extracting vital information about the world from images, but also internally representing this information so that it can be efficiently used for decisions and action [69]. But if the visual system is a computational system, what are its programs? The concept of *perceptual proxies* [24–26] has recently been proposed as a potential answer to this question. A perceptual proxy is a heuristic shortcut for how the visual system extracts data from images using simple features, such as a shape’s outline, center of mass, area, or color. The hypothesis is that, instead of computing statistics per se, the visual system relies on proxy computations across visual marks, when seeing trends in a line chart, finding maxima in a bar chart, or analyzing a distribution in a pie chart.

Perceptual proxies arise out of seminal findings on “elementary perceptual processes,” which were originally derived from a long history of empirical experiments in perceptual psychology [9, 70] and later summarized by Cleveland and McGill [2]. However, while these low-level processes can easily be applied to individual marks or groups of marks in a visualization, more composite tasks involving multiple val-

ues or general trends are more challenging to extract [71, 72]. In such situations, the visual system likely constructs proxies from these perceptual building blocks in order to support quick visual judgments.

One example is the perception of correlation in scatterplots. The perceptual process does not appear to calculate the true mathematical correlation, and there are instead proposals for multiple proxies that might underlie correlation perception [24, 73, 74], including the aspect ratio of the bounding box surrounding the points [24]. This proxy can be efficient because it relies on a rapid perceptual process of inspecting a shape boundary around the points.

Different proxies may afford not only different data patterns, but different conceptual associations of what those values might mean. The same two data points graphed as two bars or as two endpoints of a line chart can evoke different visual actions taken on visual features of the visualization. Zacks and Tversky [75] presented simple line or bar charts to participants for open description. Participants’ descriptions of bar charts overwhelmingly tended to involve discretizing words, such as “Y is higher than Z,” and descriptions of line charts entirely used continuous relations, such as “as X increases, Y decreases.” This bar-line message correspondence seems to occur because the type of mark is associated with metaphors of bars being containers or groups, in contrast to lines, which are continuous entities. Yuan et al. [25] asked participants to estimate averages in multi-value lineups of two side-by-side bar charts. Varying the number of bars in the two charts enabled them to show that the summed area of the bars is a likely perceptual proxy for the relative average value between two bar graphs.

2.4 Adversarial Visualization and Data

Central to our work is the idea to generate adversarial tasks to derive datasets, visual representations, or visual appearances that can deceive the viewer’s perception, in order to show that the viewer is taking shortcuts. One first example of such an approach in data visualization was the work by Wickham et al. [76] on graphical inference in visualization. They propose both a “Rorschach” protocol, where participants are shown essentially random data in a visualization and asked to generate insights, as well as a lineup protocol, where multiple visualizations are shown of different datasets and the task is to identify the one dataset drawn from real data.

Pandey et al. [77] studied deception in visualization by asking participants in a crowdsourced study to interpret data presented using four different distortion techniques. For each distortion type, a deceptive version, which used the technique, and a control, which did not, was used. The dataset generation in the paper was idiosyncratic and done by hand. In contrast, our adversarial dataset generation is fully automated.

The notion of “adversarial” (or “black hat”) visualizations was first proposed by Correll and Heer [78], and used the language of computer security to survey the practice of “attacks” on data visualization. Their work is largely conceptual, and only one component of their model—data manipulation—is directly relevant to our study, but the overall tenor of these ideas are consistent with our methodology.

Correll et al. [79] created crowdsourced lineups where participants saw multiple visualizations of largely “innocent” datasets with one “flawed.” They generate these

datasets using an iterative process based on three common data quality errors—spikes, gaps, and outliers—and at varying levels of data quality.

Chapter 3: Case Study: Microbiome Comparison

As a motivating example, we will describe a case study of Krona [12], the system that initially led us to investigate perceptual factors more rigorously. The scenario is the exploration of the human microbiome, or the communities of microorganisms that live in and on us. This domain is an extremely challenging one for visualization and an area of active development and interest. Since a community of organisms can be described at various levels of taxonomic granularity (i.e. genus, species, etc.), even single datasets are complex and challenging to represent. Various hierarchical techniques have been employed for the task, including Sunburst charts (as in Krona), Treemaps [80] (as in MetaTreeMap [81]), and Sankey/flow diagrams [82] (as in Pavian [83]). However, in each case, additional variables, such as change between datasets, are difficult to introduce. For scientific data, which often have control groups, the comparison of multiple data series is nonetheless critical to making sense of the underlying information. The main goal in exploring this type of data, as stated by domain experts we interviewed, is to find significant differences in the fractions of particular organisms, especially if they are pathogenic ones. Here we prototype several comparative strategies and present them to domain experts for qualitative feedback. Our goal is to see whether particular modes of comparison

affect how users interact with data and how well (qualitatively) they perform simple, but realistic tasks using actual domain data.

3.1 Method

We adapted the Krona system, which already supported animated transitions, to implement two additional comparative strategies, for a total of three (Fig. 3.1). We introduced the three techniques to two scientists studying the microbiome at the National Human Genome Research Institute in Bethesda, MD, USA. Both had prior experience with the tool for exploration of single datasets.

3.2 Task

The participants were presented with real data comparing human skin microbiomes from two time points (“M3 skin” days 0 and 1) [84]. The charts show the relative proportion of various species within each time point as well as the aggregated proportions of more general taxa. We asked the microbiologists to find significant differences between the same two time points using all three arrangements (*animated*, *adjacent*, and *mirrored*). For example, in Figure 3.1(a), *Gammaproteobacteria* (red wedges) decreases a large amount from day 0 to day 1 (left to right), but looking more specifically within this group, *Pseudomonas* actually increases. Rather than seeking a specific set of correct answers, however, we instead gathered more qualitative feedback about performing the task under the various conditions.

3.3 Results

Both participants found that animation made differences particularly salient. However, they also noted that, if the change was large, it shifted the other wedges in a disorienting way.

It was also noted by the experts that animation could be engaging for an audience when highlighting a specific difference, reiterating the findings of Robertson et al. [6]. However, both participants preferred static views when performing their own exploration or investigation. One participant preferred small multiples due to its consistency with standard sunburst charts and the ability to represent more than two samples. The other, however, preferred the mirrored split view due to the better use of space and smaller eye travel distance when making direct comparisons between constituent taxa. Additionally, the case study illuminated practical considerations of implementing these arrangements. For example, the experts pointed out that small multiples may be ideal for dissemination, which is often static and must reach a wide audience that may not be familiar with the split mirrored view.

Unsurprisingly, there was a consensus that each method had strengths and weaknesses, and would be more appropriate for specific contexts. One conclusion could be that this platform, and others, should have the flexibility to support many layouts, allowing the user to switch between them to aid the task at hand. More importantly however, we have established that efficacy of visual comparison is contextual, and how it is carried out affects interpretation of data. This will drive our subsequent investigations into which comparison methods work better for certain

tasks, and, eventually, the underlying processes that cause them to work better.

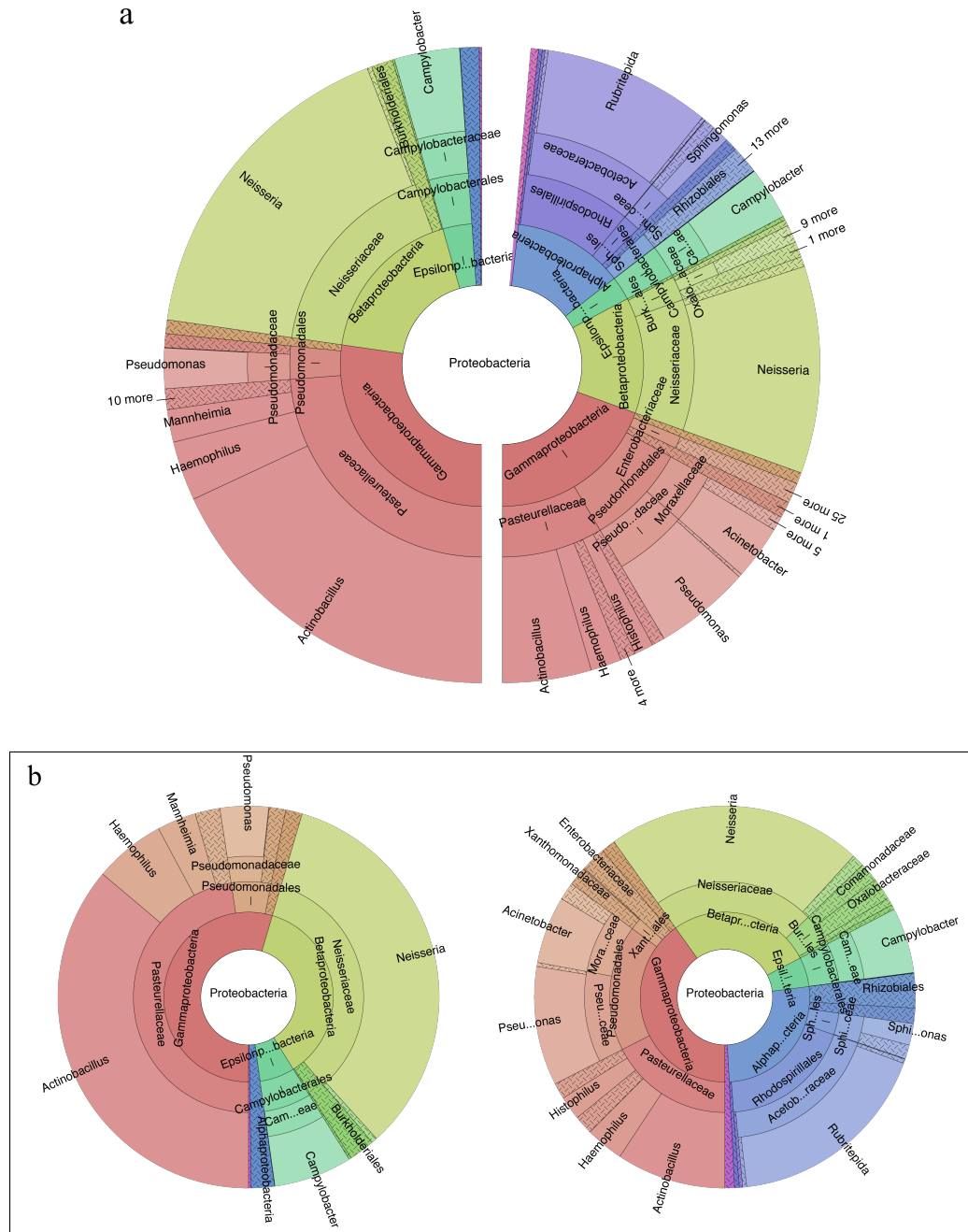


Figure 3.1: **Comparative modes implemented in Krona.** Here, the skin microbiome of an individual is represented across two time points. Higher levels (i.e. innermost rings) represent more general taxonomic categories. In (a), a single circle is split to provide mirror symmetry, corresponding to the *mirrored* arrangement in the above experiments. In (b), the standard small multiple view of the same data is shown, corresponding to the *adjacent* arrangement.

Chapter 4: User Study Framework

At the core of investigating how perceptual phenomena may impact how real people interpret charts is the user study. Paradigms for these studies can be roughly grouped into those used by perceptual psychologists and those used by data visualization designers. Perceptual studies typically investigate the atomic mechanisms of vision that are “pre-attentive,” meaning they happen without conscious direction of the mind. These studies involve simple, abstract shapes and require users to make simple, comparative judgements. Data visualization studies may require users to make more complex insights, seeking out multiple estimates in labeled data and drawing conclusions. A relatively small body of work straddles these paradigms, investigating perceptual abilities within data visualizations [2,3]. These tend to have to the simplest possible constructions that could be considered charts, containing few data points, and usually without any labels or context other than the assertion that they do, in fact represent data. Our work follows in this vein, and our experiments will be similarly constructed. Though each experiment comes with some of its own considerations, they share much of the platform, which we will describe in this chapter. Additions or deviations will be described for each task in the chapters following.

4.1 Arrangements

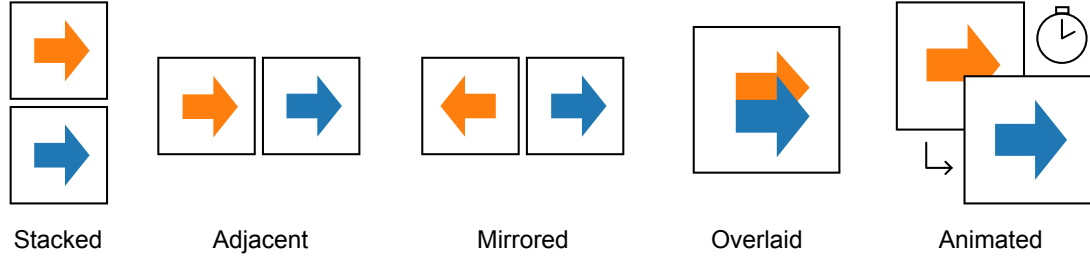


Figure 4.1: Comparative arrangement methods examined. The direction of the arrows represents the orientation of the x-axis (or, in the case of donut charts, clockwise versus counterclockwise).

Based on perceptual factors discussed above, we lay out here a set of comparative methods, which we will call *arrangements*, to probe empirically. The five arrangements we will use in our experiments are depicted in Figure 4.1.

- **Stacked:** Vertically arranged small multiples (i.e. one chart is placed above the other). Cleveland & McGill posit the aligned baselines helps judgment [2]; but this design also makes it tougher to find correspondance between paired values from each series [15, 16]. We thus include it as an expected floor to which performance of other arrangements can be compared.
- **Adjacent:** A more commonly used instance of small multiples, in which data series are placed side-by-side, allowing each pair of items to align vertically. This arrangement serves as a more realistic baseline than *stacked*.
- **Mirrored:** This “mirrored” variation of *adjacent* opposes the direction of the

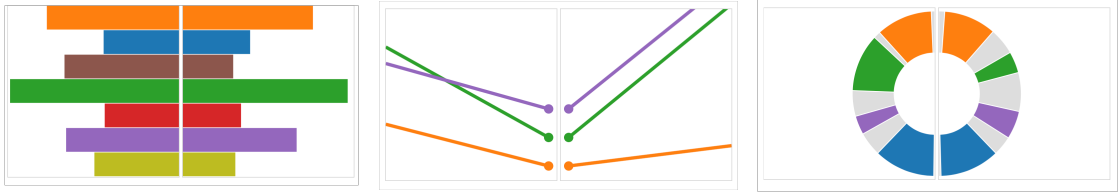


Figure 4.2: Examples of the mirrored arrangement for (left) bar charts, (middle) slope charts, and (right) donut charts.

x-axis in each chart (Fig. 4.2). For bar charts, this simply amounts to right-aligning the left chart and vice versa. For slope charts, the x-axis is reversed in the left chart, essentially negating the slope. For donut charts, we restrict each series to a semicircle. The Gestalt nature of bilateral symmetry suggests this layout could improve performance versus standard small multiples.

- **Overlaid:** A combined chart depicting both data series within the same space. Past work has claimed that overlaying values, or superposition, minimizes eye movements and memory load, and may lead to efficient comparison [21]. This technique has proven effective in a design study setting [85], but, to our knowledge, not directly confirmed empirically.
- **Animated:** In this “arrangement,”¹ a single chart is transitioned, or morphed, from one data series to another over time. As all marks transition for the same amount of time, the maximum velocity of a given mark becomes an emergent signal that directly encodes its delta. Movement is broadly processed as a primitive feature in the vision system, suggesting that this signal is potentially

¹Gleicher et al. [21] equate animation between data sets to juxtaposition across time.

beneficial for tasks in which individual items must be processed. We used cubic interpolation to ease the transitions [86], so the maximum velocity was reached at the midpoint of the impression time.

4.2 Timed Impressions

User studies in Information Visualization, such as those evaluating novel visualization methods, typically will ask users to make a judgement about a visualization and allow them to respond in their own time. The user may be instructed to be both “fast” and “accurate” in performing the task. However, this leaves open the variable of how different users may interpret this trade-off. Additionally, this would make it difficult to see effects of preattentive visual judgements. For example, the similarity of two charts can be determined by exhaustive comparison of elements, but can likely be performed much faster using the visual system’s innate ability to detect symmetry of whole shapes. In this vein, Cleveland & McGill omit tickmarks to prevent counting and instruct their participants to be “quick” [2]. Simkin & Hastie, however, explicitly limit displays to one second to control for this trade-off [9]. We follow the latter protocol, with the specific length of time determined by piloting for each experiment. After the impression, users answer as best they can given the time they had to view the charts. This allows us to evaluate performance based purely on correctness, without the time taken to respond as a confounding variable.

none of the arrangements making the task trivially easy or impossibly difficult.

To avoid the confounding factors of crowdsourcing, and the need for tedious piloting to calibrate difficulty, we thus borrow another technique for perceptual psychology: *titration*. Under this regime, rather than fixing the difficulty of the task and assessing the accuracy of responses, we aim for consistent accuracy (in this case 75%) by adjusting the difficulty of the task. The final signal is then a *titer*, which is a measure of how difficult the task had to be to reach this accuracy. To target an accuracy, adjustments are made after each response, making the task harder if the response was correct, and easier if it was incorrect. To achieve, for example, 75% accuracy, the increase in difficulty for an incorrect response is 3 times the magnitude of the decrease in difficulty for a correct response. This is termed a “staircase” from the pattern it produces (Fig. 4.3).

4.4 Dynamic Data Generation

Evaluations of information visualizations often have a fixed group of visualizations that are shown to all participants. However, if we are to dynamically titrate difficulty of tasks, it is beneficial to generate data dynamically for each trial. This ensures lack of any bias from curation of trials, and allows fine-tuning of difficulty, the freedom to test different numbers of trials, and the ability to easily replicated experiments. As all data points generated during the experiments are stored, it also provides a more diverse source for downstream analyses. Dynamic data generation, is not without its challenges, though, chief among them being the need to ensure

there are not emergent signals in the absence of curation. For example, if we generate data sets from distributions with two means, the one with the higher mean is more likely to contain the largest point overall, which could allow participants to take “shortcuts,” and this must be corrected for algorithmically. Each task comes with its own such considerations, as will be discussed in the coming chapters.

4.5 Rendering

Charts were rendered in the participant’s web browser in real time using the D3 [87] JavaScript library. Size of the charts, in pixels, varied by experiment. Note, however, that the actual number of screen elements corresponding to a “pixel” can vary with hardware configuration, due to the advent of HDPI (high dot-per-inch) displays. Bar charts would not be affected by this variable because of their orthogonal nature, and we chose sufficient line thickness to mitigate the effect for slope charts. All charts were drawn on white backgrounds, with faint gray boundaries delimiting the chart areas. As we are investigating elementary visual operations, similarly to prior studies [2, 3], we omit tickmarks, as they may encourage participants to count rather than judge. The web page automatically initiated full-screen browsing mode to avoid distraction during the study, though the persistence of this state was not enforced programmatically.

4.6 Crowdsourcing

As is common for modern user studies, we used compensated crowdsourcing, in this case via Amazon Mechanical Turk, which allowed dozens of participants to participate in each experiment in a matter of days. While this method comes with its own issues, for example heterogeneity of experimental conditions and reliability of participants, it has been shown that perceptual results can be faithfully reproduced in this setting, provided the proper care is taken [3].

4.6.1 Training

Before training, participants were shown examples of stimuli and the task. Before each arrangement block of trials, participants were given a time-unconstrained version of the task, which they were required to answer correctly before proceeding. Additionally, the first non-animated arrangement given to a participant followed untimed training with 3 timed training trials, which were identical to the real trials except that they always had the easiest (largest) titer. Data were regenerated on incorrectly answered training answers to minimize answering by elimination.

4.6.2 Participant Recruitment and Payment

We recruited non-expert participants through the Amazon Mechanical Turk Platform. We limited participation to the adults in the United States due to tax and compensation restrictions imposed by our IRB. Only workers with a 95% approval rate on the platform were eligible. We also screened participants to ensure at

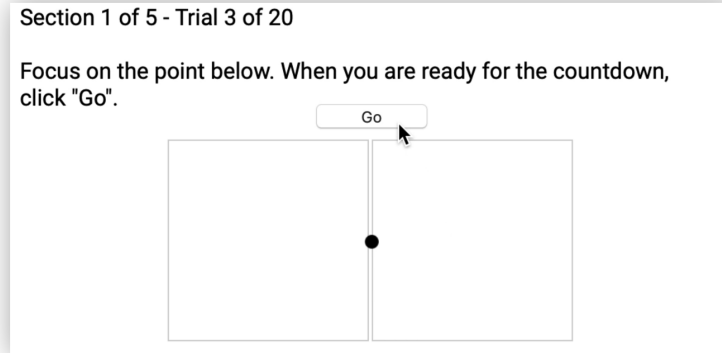


Figure 4.4: To ensure a participant was ready and focusing, a screen like this one was shown before each trial, followed by a countdown and then the impression.

least a working knowledge of English; this was required to follow the instructions in our testing platform. Based on expected task completion times, participants were compensated at a rate consistent with an hourly wage of \$8/hour (the U.S. federal minimum wage in 2020 is \$7.25). Participants were asked to self-select out of the study if they had color vision deficiencies. Worker IDs were used to ensure uniqueness of participants across all such combinations. Based on power analyses from initial pilots, we recruited at least 50 new participants for each experiment. A total of 435 workers were recruited for participation in the experiments.

4.7 Procedure

Before each trial began, the screen contained a centrally placed fixation dot and outlines of where the charts would appear (Fig. 4.4). Participants clicked a button to start the trial. After a countdown, the visualization appeared for a short,

fixed time. At the end of the impression, a prompt for response was provided, either by removing one of the data series and making the remaining one clickable or by removing all charts and providing color-coded buttons. Participants were informed if they were correct and, if incorrect, what the correct answer was. This feedback was provided to make the task more engaging and to reinforce the goal. Between trials, the titer was adjusted based on the response (if incorrect, the titer was made larger for the next trial; if correct, the titer was made smaller). Each participant completed one experiment, each with all arrangements, which were blocked. The order of the arrangement blocks was changed between participants by cycling through all rotational permutations of the base sequence [stacked, adjacent, mirrored, overlaid, animated] and all rotational permutations of the reverse of this sequence..

Chapter 5: Experiment 1: Maximum Delta Task

In our first experiment we task participants with finding the maximum delta for two data series. In other words, from one series to the next, which data point changed the most? This could be an increase or decrease, defined by absolute change, as opposed to percent change. Difficulty is increased by reducing the largest delta while increasing distractor deltas, so the maximum is less distinguishable. A bimodal distribution of absolute values decouples the largest delta from the largest or smallest absolute value in any single set.

5.1 Experimental Setup

Since the choice of visual encoding channel could interact with the choice of arrangement, for this first experiment we evaluated several encodings. To ensure each chart type provided an appropriate range of difficulty, parameters such as the number of data points had to be adjusted. These parameters were determined during internal piloting, resulting in the following configurations:

- **Bar charts:** Standard charts in which the length corresponds to the datum.

Each series contains 7 data points.

- **Slope charts:** Simplified line charts with just two points in each line, (0 and

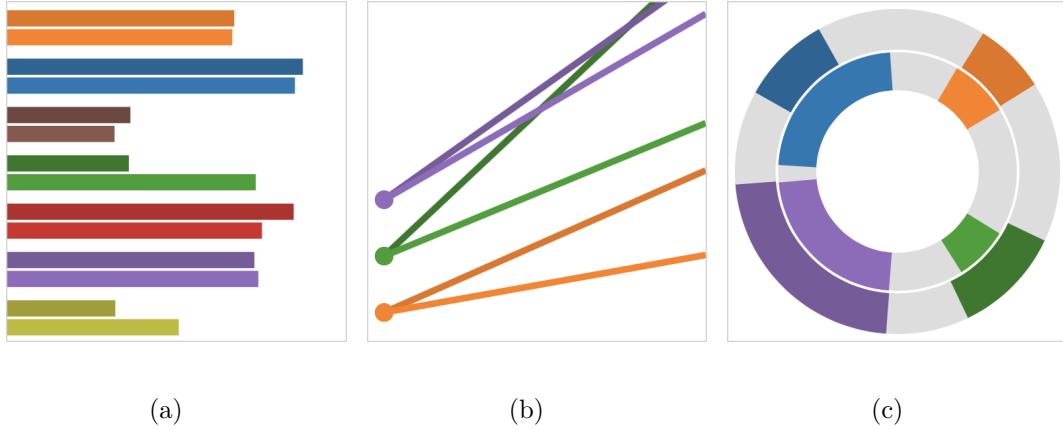


Figure 5.1: Encodings used for the Maximum Delta task, shown here for the overlaid arrangement: (a) bar, (b) slope, (c) donut.

a generated datum), reducing them to slopes. Each series contains 3 data points.

- **Donut charts:** Rings in which the data are represented by angular sector. For the purposes of experimental control, they differ from standard donut charts in several ways: (i) gray distractors are used as buffers to allow adjacent data to change size while remaining in the same position, (ii) overlaid arrangements, which are non-standard for donut charts were implemented with concentric rings, aligning the centers of corresponding colors, and (iii) mirrored arrangements were implemented by limiting each chart to 180 degrees, allowing the two series to form a complete circle. Each series had 4 data points.

Each individual chart (that is, for a single data series), had a square dimension of 256 pixels for all trials. Subsets of the Tableau 10 [88] were chosen to maximize (qualitatively) perceived uniqueness; 7 for bars, 3 for slopes, and 4 for donuts. For the overlay arrangement, the saturation and luminance of each color were slightly

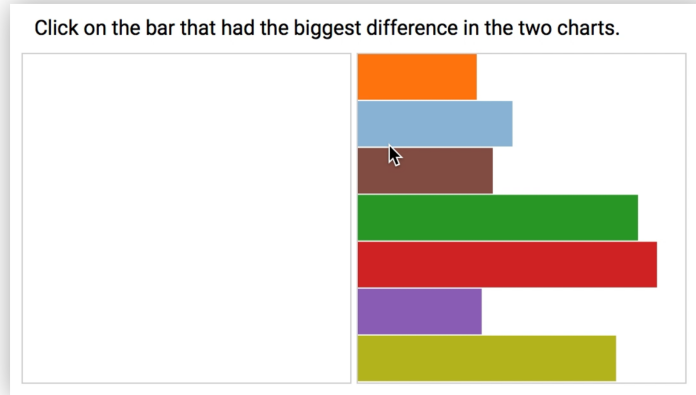


Figure 5.2: The response prompt for the Maximum Delta task. One dataset was removed, and the bars themselves (or other encodings) were used as buttons.

reduced in one dataset to distinguish adjacent elements. Other arrangements kept the original colors consistently across pairs of data sets. Note that, since donut charts have no explicit orientation, we omitted the *stacked* arrangement for this encoding because of redundancy with *adjacent*. Each participant completed all arrangements (4 for donut, 5 for others) for a single stimulus type (bar, slope, donut). There were twenty trials for each arrangement of bars and slopes, and thirty for donuts, based on power analysis from piloting. The impression time for both static charts and animation was 1.5 seconds. Following the impression, one data series was removed, leaving all colors of the second series to act as response buttons so that a participant could simply click on any of the bars, lines, or arcs, in the remaining series (Fig. 5.2).

5.2 Data Generation

A pair of datasets with controlled deltas was generated by varying points of one dataset to create another. However, simply increasing or decreasing one data point more than others—out of, say, of a normal distribution—would make it much more likely to be the largest or smallest, circumventing the task. It was thus necessary for proper evaluation of the task to devise a novel data generation algorithm. Our method creates a bimodal distribution corresponding to the two extremes of a chosen maximum delta, ensuring that these points are well masked by other data. The magnitude of this delta, and thus the difficulty of the task, is controlled parametrically by the titer value provided to the generation algorithm. In addition to changing the maximum, changing the titer also changes deltas of distractors. At the minimum (smallest difference) titer, every data point is changed a small, equal amount (note that it is, by design, impossible to do better than chance at this level, and in practice it is never reached). At the maximum (largest difference) titer, there are only two possible values for the data points—the maximum uses both, while the others stay at one and do not change at all. The data generation routine is depicted at a high level in Algorithm 1. In summary, for a given titer value t , the biggest mover will change by t times the chart’s range (from minimum value to maximum value). The biggest moving distractor will change by $1 - t$ of that, the next biggest moving distractor $1 - t$ of the first distractor, and so on. For example, at a titer of 0.75, the delta of the biggest mover will cover $3/4$ the full range of the chart, the delta of the first (randomly placed) distractor will cover $0.75 \times 0.25 = 3/16$, and

that of the next will cover $0.75 \times 0.25 \times 0.25 = 3/64$. The outputs of this algorithm were linearly transformed as appropriate for the stimuli, e.g. to add minimum width to bars. Though higher titers should always be easier, in practice, we found that difficulty increased above 0.75 due to alignment of bars. We thus capped the titer at 0.75 to prevent participants from getting stuck in a valley of (ostensibly) low difficulty. We confirmed the regularity of the data before the experiment by running multiple iterations of the data generation routine and observing the ordinal ranking of the answers among the distractors. While there do appear to be areas of bias, we deem it highly unlikely that detection of these patterns would be easier for a participant than performing the task as intended.

5.3 Hypotheses

We expect the overlaid arrangement to serve as a ceiling for performance in the context of this task because of the close proximity of corresponding elements. We also expect that, among small-multiple arrangements, mirror might perform best because it could allow the vision system’s preattentive identification of symmetry to make the biggest mover appear as an outlier. Between horizontal juxtaposition (“adjacent”) and vertical (“stacked”), however, expectations are less clear. Adjacent aligns corresponding bars (of the same color) vertically, which would make it easier to observe both. However, stacked aligns their baselines horizontally, which prior studies have indicated facilitates quantitative comparisons.

Algorithm 1 Max-delta data generation

```
1: procedure MAXDELTA( $c, t$ ) ▷  $c$ :=cardinality,  $t$ :=titer
2:    $a \leftarrow [], b \leftarrow []$ 
3:   for  $i = 0$  to  $c - 1$  do
4:      $r \leftarrow rand()$  ▷  $r \sim U, r \in \mathbb{R}, 0 \leq r \leq 1$ 
5:      $x \leftarrow t \cdot \sqrt{\frac{r}{c-i}}$ 
6:      $y \leftarrow x + t(1 - t)^i$ 
7:     if  $i \% 2 == 1$  then
8:        $x \leftarrow 1 - x$ 
9:        $y \leftarrow 1 - y$ 
10:    if  $rand() < 0.5$  then
11:      push  $a, x$ 
12:      push  $b, y$ 
13:    else
14:      push  $a, y$ 
15:      push  $b, x$ 
16:  return  $a, b$ 
```

5.4 Results

To evaluate whether arrangement affected the precision with which participants could identify the maximum delta, we computed each observer’s mean titer values from the final 5 trials for each arrangement. Titrers are inversely related to difficulty: smaller titers for a chart arrangement indicate that subtler, rather than larger, differences were required to elicit a mixture of correct and incorrect responses. Based on the outlier criteria described in Seciton 4.6.2, 3 participants were excluded from experiments with bar encodings, 4 from slopes, and 2 from donuts.

5.4.1 Exp. 1A: Bar charts

Exp. 1A: Bar charts (max delta task)

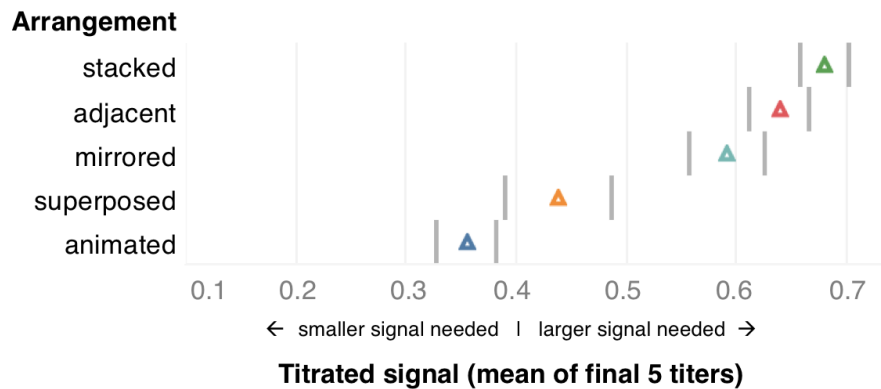


Figure 5.3: Mean of final 5 titer values across participants performing the Maximum Delta task with bar charts. Gray bars represent 95% confidence intervals.

Figure 5.3 displays the mean final 5 titer values for Experiment 1A. In bar charts, two patterns in participant titer values were striking. First, the Animated bars outperformed bars that were Overlaid and all other arrangements. Second,

Exp. 1B: Slope charts (max delta task)

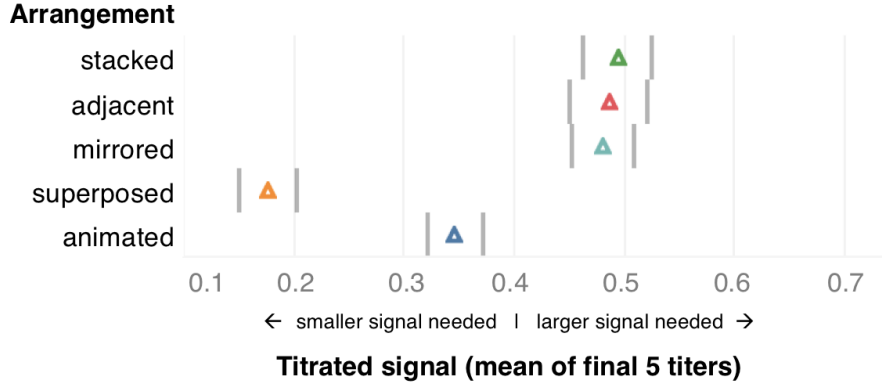


Figure 5.4: Mean of final 5 titer values across participants performing the Maximum Delta task with slope charts. Gray bars represent 95% confidence intervals.

within Small Multiples, a Mirrored arrangement is better than a Horizontal or Vertical one.

These observations were validated in a within-subjects ANOVA. Final titer values for bar charts were affected by arrangement, $F(2.98, 137.23) = 103.23$, $p < .001$, $\eta_p^2 = 0.69$, Greenhouse-Geisser corrected for violations of sphericity. Planned comparisons assessed pairwise differences between arrangement types. Titers for animated bars were significantly more precise than those that were overlaid, $t(46) = 3.42$, $p = .001$. Participants also achieved more precise titer values with horizontally mirrored small multiples compared to non-mirrored small multiples that were horizontally arranged, $t(46) = 2.73$, $p = .009$, and vertically arranged, $t(46) = 4.82$, $p < .001$.

Exp. 1C: Donut charts (max delta task)

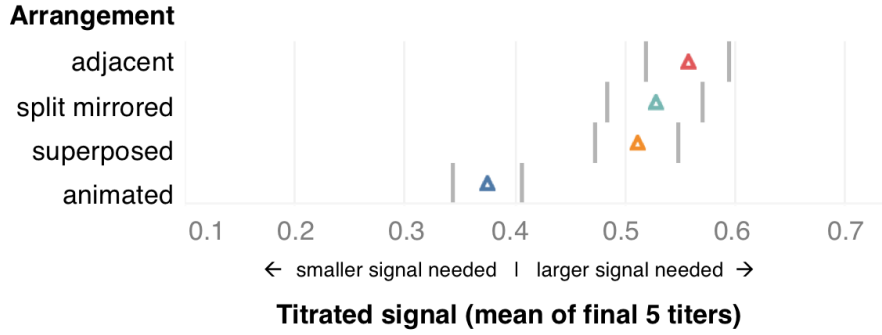


Figure 5.5: Mean of final 5 titer values across participants performing the Maximum Delta task with donut charts. Gray bars represent 95% confidence intervals.

5.4.2 Exp. 1A Floor Effect

For the final five trials, accuracy was low for stacked (57%), adjacent (61%), and mirrored (64%) arrangements, with large titers near the maximum titer of 0.75 (0.68, 0.64, and 0.59, respectively). By comparison, for animated arrangements, accuracy was 74.6% and the mean titer was 0.35. Participants reached the maximum titer on 28% of stacked trials and 15% of adjacent trials. By comparison, the maximum titer was reached on 6% of mirrored trials, 5% of overlaid trials, and 0% of animated trials. The histograms in Figure 5.6 illustrate titer distributions for all trials for each arrangement. These floor effects suggest that for stacked and adjacent charts, subjects reached the artificial floor (max titer) and continued making errors without subsequent adjustments to the titer value, such that their final titer value reflects not their ability to do the task but the capped titer value. As such, Experiment 1A is not able to quantify the true floor of performance for these arrangements.

Note that this floor issue is unavoidable for many tasks. One solution for future research is a longer display time, but that could make more effective arrangements (e.g. overlaid) too easy, resulting in a ceiling effect and preventing comparison. Another solution is to conduct secondary tests of arrangements that are close in performance, using combinations of titer ranges and timings that best drive apart performance.

In summary, although this data set cannot be appropriately used to directly compare the mean titers between stacked and adjacent arrangements, it is clear that the MAXDELTA task was highly difficult in stacked and adjacent bar charts.

There was no evidence for floor effects in subsequent experiments.

5.4.3 Exp. 1B: Slope charts

In slope charts, titer values were generally more precise and there were slightly different observations as a function of arrangement. First, Overlaid slopes outperformed all other arrangements (including Animated). Second, different types of Small Multiple arrangements did not yield differing titer values (Fig. 5.4).

These observations were validated in a within-subjects ANOVA. Final titer values for slope charts were affected by arrangement, $F(4, 180) = 101.87$, $p < .001$, $\eta_p^2 = 0.69$. Titer histograms did not indicate floor effects. Planned comparisons assessed pairwise differences between arrangement types. Titers for overlaid slopes were significantly more precise than those that were animated, $t(45) = 10.13$, $p < .001$. There was no evidence that participants achieved more precise titer values with

Titer histograms for all analyzed observers in Exp. 1A

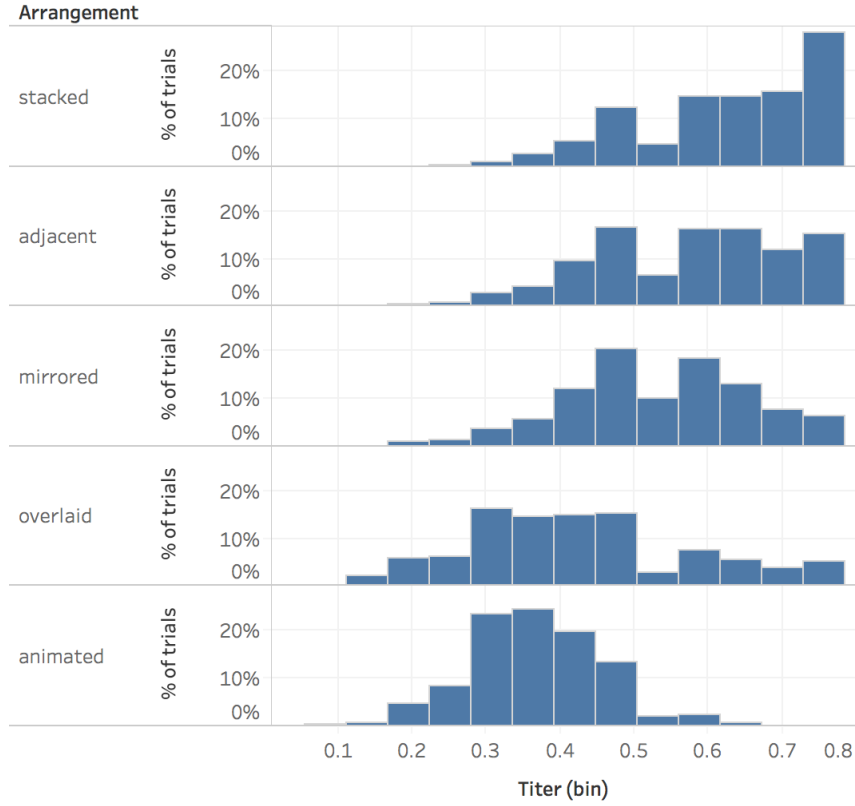


Figure 5.6: Histograms of titers (across all trials) by arrangements from Experiment 1A, for all non-excluded participants. Participants disproportionately reached the maximum titer value (0.75) for stacked (vertical small multiple) and adjacent (horizontal small multiple) arrangements.

horizontally mirrored small multiples compared to non-mirrored small multiples that were horizontally arranged, $t(45) = .25$, $p = .8$, or vertically arranged, $t(45) = .77$, $p = .45$. Accuracy exhibited similar patterns as titer values.

5.4.4 Exp. 1C: Donut charts

The mean final 5 titer values for donut charts were affected by arrangement, $F(3, 141) = 22.96$, $p < .001$, $\eta_p^2 = 0.33$ (Fig. 5.5). Titer histograms did not indicate

floor effects. Animated donuts outperformed all other arrangements for the max-delta task. There was no evidence that the split mirrored arrangement outperformed the horizontal small-multiple donuts, $t(47) = 1.26$, $p = .21$. Accuracy exhibited similar patterns of titer values.

5.5 Discussion

For the Maximum Delta task, animated charts consistently outperformed all small multiple arrangements. Findings were mixed for overlaid visualizations: they outperformed all other arrangements (including motion) for slope charts, were better than any arrangement of multiple bar charts, and did not seem to confer strong benefits over small multiple arrangements for donuts. Finally, mirrored small multiple arrangements marginally allowed participants to better identify the max-delta series (compared to other horizontal arrangements) only in bar charts. Although animated charts outperformed others for the goal of the task, and as such is useful if an analyst's goal is to rapidly identify individual data points with the largest improvement or impairment, it might not be an optimal encoding for other goals of the observer or designer. Specifically, a maximum delta task may be a special case in which velocity information directly encodes individual data deltas but does not directly encode the visual information that observers use to inform other judgments, such as the overall correlation or mean. As an example of a caveat of extending the lessons of perceptual studies to a more ecological valid environment, case study participants noted disorientation when wedges were animated in Krona (see Sec. [3.3](#).

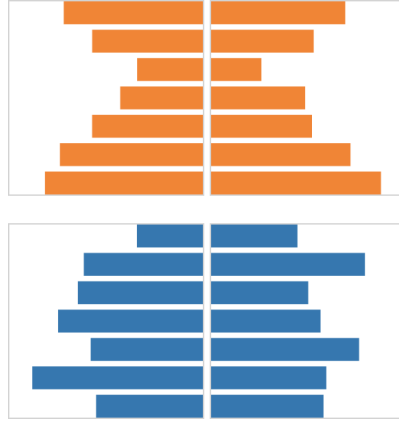
This is because, unlike in those studies, the positioning of the wedges could not be controlled using distractors. However, it also could suggest work to be done to take advantage of the benefits of animation seen in these studies—for example, perhaps wedge ordering could be optimized in animated sunburst plots to minimize offsetting during a transition, as has been done for stability in animated Treemaps [\[89\]](#).

Chapter 6: Experiment 2: Correlation task

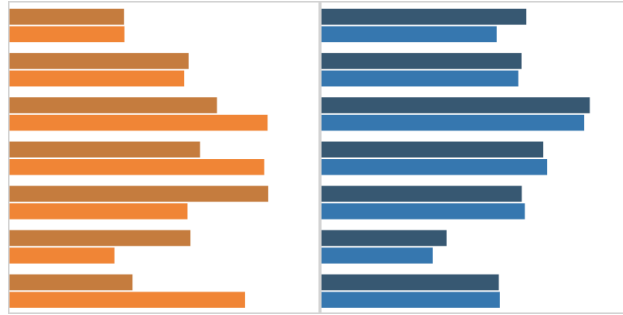
For our second experiment, we chose a more holistic task to contrast the individual nature of the Maximum Delta task: out of *two pairs* of charts, which pair exhibits the most correlation between its two series? We base the methodology on past studies for similar tasks [73, 74]. Difficulty of this task is adjusted by varying the correlation of the target series pair, while leaving the control pair at a low, fixed correlation. Since correlation may be too esoteric of a concept for crowdsourcing, we instructed participants to choose the “most similar pair” and ensured that each chart in a pair had comparable means and standard deviations.

6.1 Experimental Setup

For this experiment (and all subsequent experiments), in order to focus our attention and resources, we used only bar charts, which had the most interesting results in Experiment 1. Since in total four data series are required for the comparison, four charts are rendered for all arrangements except overlaid, which has two charts, each with two data series (Fig. 6.1. Renderings with four charts used a square dimension of 200 pixels, while renderings with two charts used a square dimension of 141 pixels (producing equivalent total chart area). Static charts were



(a)



(b)

Figure 6.1: Example renderings of the Correlation task, shown for (a) mirror and (b) overlaid arrangements.

shown for 3 seconds, to account for the doubled number of charts, while animation remained at 1.5 seconds to preserve velocity. At the end of the impression, one data series from each pair was removed (always the leftmost or uppermost, as an arbitrary convention) to obscure the true similarities, such that one chart each for orange and blue colors remained as response buttons.

6.2 Data Generation

Randomized pairs of series with given correlations were created using simulated annealing in an algorithm inspired by Matejka and Fitzmaurice [90]. Means and standard deviations were fixed within 10 percent of the range to ensure correlation was analogous to “similarity”, as described in the instructions. Correlation between the series was calculated using Pearson’s correlation coefficient and transformed according to the optimal formula for perceptual estimation according to Rensink & Baldridge [74]. Titters we report for this experiment thus correspond to $g(r)$ in Equation 7 of the latter study.

6.3 Results

Exp. 2: Bar charts (correlation task)

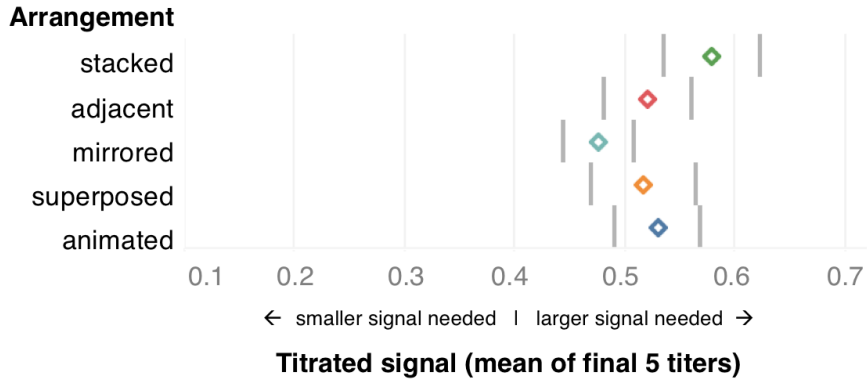


Figure 6.2: Mean of final 5 titer values across participants performing the Correlation task (with bar charts). Gray bars represent 95% confidence intervals.

The mean final 5 titer values for Experiment 2 were affected by arrangement, $F(3.22, 144.95) = 6.50$, $p < .001$, $\eta_p^2 = 0.13$, with no indication of floor effects

(Fig. 6.2).

In contrast to Experiment 1, it is apparent there was no benefit of animation over other arrangements: participants struggled to use motion to extract and compare correlations between data sets. Observer performance had resulted in staircasing of the mean correlation (Pearson’s R) to 0.74 for observers to reliably choose it over the base pair correlation of 0.20.

We conducted planned comparisons to assess whether mirrored small multiples yielded more precise titers than the other small multiple arrangements. Participants achieved more precise titer values with mirrored compared to adjacent arrangements, $t(45) = 2.13, p = .04$. They were able to perform correlation comparison when the target correlation was 0.70 in mirrored charts, but needed a correlation of 0.75 for the same performance in adjacent chart arrangements. Adjacent bar charts outperformed stacked ones, $t(45) = 3.31, p = .002$, such that for these trials the correlation of the correct pair was 0.82 for stacked charts.

Accuracy exhibited largely similar patterns as titer values, with the exception that there was only marginally significantly higher accuracy for mirrored compared to adjacent charts, $t(45) = 1.95, p = .057$.

6.4 Summary

For the Correlation task, animation did not provide the benefits of the Maximum Delta task. Instead, mirrored bar charts outperformed all other arrangements for detection of correlated data.

Chapter 7: Experiment 3: Maximum Mean Task

Continuing to investigate holistic comparisons, we ask: of two sets, which had the largest average (mean) value? Difficulty is increased by reducing the delta between the mean values, so that the difference between sets is less distinguishable. Displays were controlled so that the largest single-item in a chart was not predictive of that chart having the largest mean and so that charts in a trial were of approximately equal variance. Within-chart variance ranged from .04 to .09. Harder discriminations (smaller mean deltas) spanned the low to high variance range, whereas easier discriminations tended to be lower variance.

7.1 Experimental Setup

In contrast to previous experiments, at the end of the impression, both sets of data were removed from the display (in both Experiments 1 and 2, the data from each answer spanned multiple charts, such that the data in one chart could remain for response; this is not the case here). Participants then clicked on the orange or blue button corresponding to the orange or blue set of bars to provide a response (Fig. 7.1). The instruction given was to “Click on the chart that had the biggest mean values.” Based on previous work, we predicted $N = 50$ would provide



Click on the set with the highest
mean:  or 

Figure 7.1: The response prompt for the Maximum Mean task. Unlike Experiments 1 and 2, both datasets were removed, and color-coded buttons were used.

sufficient statistical power to reliably detect the presence or absence of an effect of arrangement.

7.2 Data Generation

Data generation for both this task and the following task (Maximum Range) was based on a bounded distribution function (Alg. 2), which samples from a normal distribution but ensures that all values lie within a given range, in addition to the mean falling within some tolerance of a target. The procedure for generating data for the Maximum Mean task based on bounded distributions is described in Algorithm 3. The two extreme values that bounded data generation were directly included in a randomly selected chart, ensuring that the highest or lowest individual value did not correlate with the correct answer.

Algorithm 2 Bounded distribution

```
1: procedure BOUNDEDIST( $\mu, \sigma, min, max, ext = false$ )           ▷ ext:=include
    extrema

2:    $\tau \leftarrow 0.005$                                            ▷ tolerance

3:   if ext then

4:      $a \leftarrow [min, max]$ 

5:   else

6:      $a \leftarrow []$ 

7:   while length( $a$ ) <  $c$  do

8:      $r \leftarrow norm(\mu, \sigma)$                                ▷  $r \sim N(\mu, \sigma)$ 

9:     if  $r \geq min$  &&  $r \leq max$  then push  $a, r$ 

10:  while abs( $\mu - \mu'$ ) >  $\tau$  do

11:     $\Delta \leftarrow \mathbf{rand}() * (\mu - \mu')$                        ▷  $\mathbf{rand}() \sim U(0, 1)$ 

12:     $i \leftarrow \mathbf{randInt}(2, c - 1)$ 

13:    if  $max \leq a_i + \Delta \leq max$  then

14:       $a_i \leftarrow a_i + \Delta$ 

15:       $\mu' \leftarrow \mathbf{mean}(a)$ 

16:  return shuffle( $a$ )
```

Algorithm 3 MaxMean data generation

```
1: procedure MAXMEAN( $c, t$ ) ▷  $c$ :=cardinality,  $t$ :=titer  
  
2:    $\sigma \leftarrow \frac{3}{8} * t$   
  
3:    $ext \leftarrow \mathbf{randInt}(0, 1)$  ▷ which array to include extrema in  
  
4:    $a \leftarrow \mathbf{BoundedDist}(\frac{5}{8} - t * \frac{3}{16}, \sigma, \frac{1}{4}, 1, ext == 0)$   
  
5:    $b \leftarrow \mathbf{BoundedDist}(\frac{5}{8} + t * \frac{3}{16}, \sigma, \frac{1}{4}, 1, ext == 1)$   
  
6:   if  $\mathbf{mean}(a) > \mathbf{mean}(b)$  then  
  
7:      $\mathbf{swap}(a, b)$  ▷ ensure correct answer at low titers  
  
8:   return  $a, b$ 
```

7.3 Results

We computed each observer’s mean titer values from the final 10 trials for each arrangement. We used the final 10 trials because visual evaluation of trial-by-trial data suggested that this was approximately when the staircase procedure stabilized around a narrow range of titers, for most participants. Thus we analyze the final 10 titer values achieved for each of the five arrangements, for each subject. We excluded one participant based on the criterion described in Section 4.6.2. We also adopted a second criterion for this experiment. In a staircase procedure, the goal is to find a converged titer value for which a participant is 75% accurate. The procedure fails if a participant repeatedly reaches ceiling performance (a minimum titer value of 0.01) or floor performance (the maximum titer value of 1.0) because at this point the stimuli cannot titrate difficulty beyond these floors and ceilings. Because viewers performed tasks for 5 arrangements, we excluded participants for whom there were

at least 5 trials of floor or ceiling titer values. These criteria excluded 0 from the MAXMEAN task, but for MAXRANGE there was 1 trial in which a participant reached ceiling performance and 109 trials who repeatedly reached the floor titer (largest delta). We excluded 7 participants for whom there were at least 5 (up to 22) trials of floor titer values (one of whom was also the participant excluded with the standard deviation procedure), leaving $N = 49$ for the MAXMEAN task and $N = 47$ for MAXRANGE. Figure 7.2 displays the mean final 10 delta values for this task. Means could be discriminated when they differed by approximately 5-8% of the chart axis, and the precision of visual comparison was affected by arrangement. Precision was better in stacked relative to adjacent charts for the Maximum Mean task, $t(49) = 2.73$, $p = .009$. Superposed charts resulted in the lowest precision for this task.

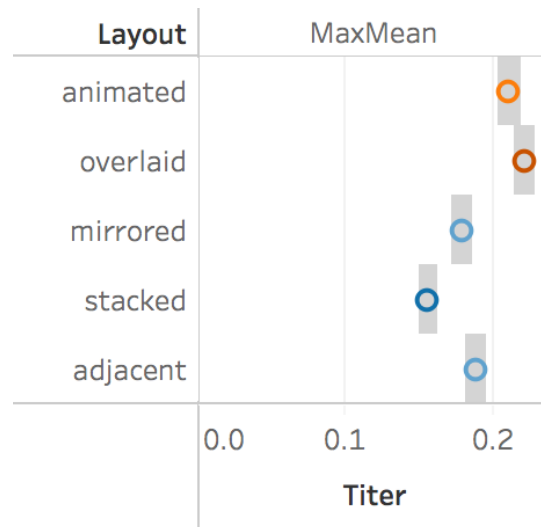


Figure 7.2: Means of averaged final titer values across participants performing the Maximum Mean task. Smaller titers correspond to more precise differences between means (range widths). The precision was affected by chart arrangements. Gray bars represent 95% confidence intervals.

The goal of a staircase procedure is to titrate the task’s difficulty so difficulty might change across arrangements, but that accuracy is equivalent between arrangements. Mean accuracy in the task for each arrangement ranged from 76.4% (stacked) to 79.9% (mirror), with no evidence that accuracy was different between arrangements. This suggests the staircase procedure reliably converged for this task.

7.4 Summary

Since the Maximum Mean task requires judgement of entire data series, somewhat like the Correlation task, one might conjecture that it would yield a similar pattern of arrangement-based performance. However, this is far from what we find in our experiments. Stacked charts, which performed the worst for Correlation, instead perform the best.

Chapter 8: Experiment 4: Maximum Range Task

For our last experiment, we chose a task that involves individual item comparisons, like the Maximum Delta task, but within each data series rather than across them: of two sets, which had the widest range between its min and max values? Difficulty is adjusted by varying the delta value between the range widths of the two charts. Since range may be a less widely-understood concept, we gave our participants a detailed description with a simple example, both at the start of the trials and each time they were incorrect in training trials. Since we expected that more participants would struggle to understand the MaxRange task, we collected data from 54 workers for this task.

8.1 Experimental Setup

Like the Maximum Mean task, both sets of data were removed from the display and participants then clicked on the orange or blue button corresponding to the orange or blue set of bars to provide a response. The instruction given was to “Click on the chart that had the widest range between min and max values.”

8.2 Data Generation

Like the Maximum Mean task, the Maximum Range data generation procedure is based on sampling from a bounded normal distribution (Alg. 2), except this time constraining the bounds further to create series with particular ranges. The procedure is described in Algorithm 4. The main concern for visual shortcuts for this task is that series generated with wider ranges are more likely to have the smallest or largest overall value. This is accounted for by abutting one range to the left chart bound and one to the right, such that, when ranges are small, one chart will have only short bars and one will have only long bars. Which one of these is the chart with the widest range is chosen randomly, ensuring that a participant can neither simply choose the chart with the shortest bar nor the longest bar and perform better than chance.

8.3 Results

As for the Maximum Mean task, we used the mean of the final 10 titer values as the signal of the precision of comparisons for a given arrangement. Figure 8.1 displays the mean final 10 delta values for the Maximum Range task. These titer values correspond to the differences between the charts being compared. Range widths could be discriminated when they differed by approximately 14-17%. As in previous tasks, the precision of visual comparison was affected by arrangement.

Titer values for the present experiment were analyzed with a mixed ANOVA

Algorithm 4 MaxRange data generation

```
1: procedure MAXRANGE( $c, t$ ) ▷  $c$ :=cardinality,  $t$ :=titer  
  
2:    $\sigma \leftarrow \frac{3}{8} * (t + 1)$   
  
3:    $t' \leftarrow 1 - t$   
  
4:    $flip \leftarrow \text{randBool}()$   
  
5:   if flip then  
  
6:      $min \leftarrow \frac{1}{4} + \frac{1}{4}t'$   
  
7:      $max \leftarrow 1 - \frac{1}{2}t$   
  
8:      $a \leftarrow \text{BoundedDist}(\frac{1}{2}(min + max), \sigma, min, max, true)$   
  
9:      $b \leftarrow \text{BoundedDist}(\frac{1}{2}(min + 1), \sigma, min, 1, true)$   
  
10:  else  
  
11:     $min \leftarrow \frac{1}{4} + \frac{1}{2}t$   
  
12:     $max \leftarrow 1 - \frac{1}{4}t'$   
  
13:     $a \leftarrow \text{BoundedDist}(\frac{1}{2}(min + max), \sigma, min, max, true)$   
  
14:     $b \leftarrow \text{BoundedDist}(\frac{1}{2}(\frac{1}{4} + max), \sigma, \frac{1}{4}, max, true)$   
  
15:  return  $a, b$ 
```

to test for experiment-level and arrangement-level effects.

Titer values varied between experiments, $F(1, 94) = 9.06$, $p = .003$, $\eta_p^2 = 0.09$, but this is likely because the titer values scale to different stimulus changes between the two experiments. As such we avoid a meaningful comparison between differing titer values.

Mean accuracy ranged from 75.2% (superposed) to 84.7% (stacked), and a repeated measures ANOVA found that accuracy consistently differed between arrangements, $F(4, 184) = 4.34$, $p = .002$. The staircase procedure did not reliably converge for all arrangements in the task due to large effects of arrangements on people’s ability to perceive range widths. Stacked charts allowed for higher accuracy and high precision than other arrangements.

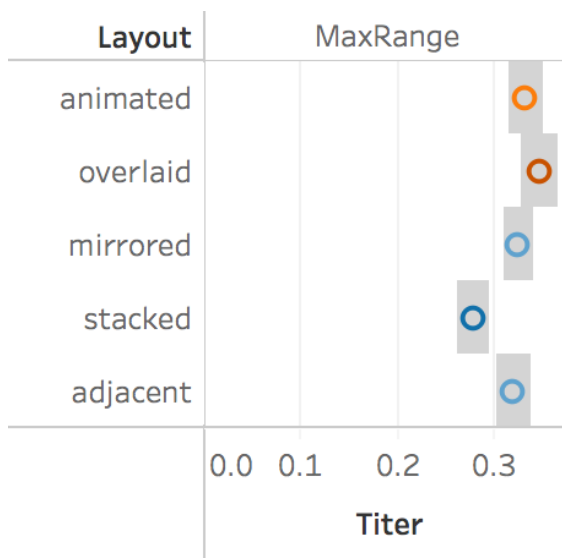


Figure 8.1: Means of averaged final titer values across participants performing the Maximum Range task. Smaller titers correspond to more precise differences between means (range widths). The precision of the task was affected by chart arrangements. Gray bars represent 95% confidence intervals.

More meaningful is that there was a significant effect of arrangement on precision, $F(3.09, 290.7) = 8.17$, $p < .0001$, $\eta_p^2 = 0.08$, without evidence for an interaction between arrangement and experiment, $F(3.09, 290.7) = .34$, $p = .85$, $\eta_p^2 = 0.004$, both Greenhouse-Geisser corrected. This suggests that arrangement produces largely similar effects on the precision of visual comparisons of means and of ranges.

Precision in stacked charts, relative to adjacent charts, was not significantly better in the Maximum Range task, $t(47) = 1.70$, $p = .09$. Overlaid charts resulted in the lowest precision for this task, as for for Maximum Mean. Note that these patterns are strikingly different compared to prior evaluation of visual comparisons of items, which were best supported by animated and superposed charts.

8.4 Summary

Like the Maximum Mean task, the Maximum Range task requires the extraction of a summary value for each data series. However, in this case, comparison of individual items within each series is more important. This distinction does not seem to drastically affect the overall ranking of arrangements. As with Maximum Mean, stacked charts perform best, which remains surprising given their poor performance in the first two tasks (Maximum Delta and Correlation). Further, this task does not seem to share any similarity with the other item-based task (Maximum Delta), with an arrangement performance profile that is almost opposite: not only is stacked at the top for Maximum Mean rather than the bottom, overlaid and

animated are at the bottom, rather than the top.

Chapter 9: Perceptual Proxies

The precision of visual comparison of Maximum Mean and Maximum Range tasks were best supported by vertically stacked charts, and least supported by superimposed charts. This is in contrast to Maximum Delta, which was best supported by animated and superimposed charts, and Correlation, which was best supported by mirrored charts. Thus, unlike early experiments on elementary perceptual encodings, no clear, universal ranking emerges for comparative arrangements. This suggests that the vision system is not simply extracting individual values from elementary encodings and computing summary statistics on those values, as, if that were the case, one would not expect arrangement to have such a large impact. This begs the question of what the brain is actually doing, if not computing these statistics.

A relatively new concept in vision science [25], a *perceptual proxy* is a visual shortcut based on a spatial feature of a visualization that could conceivably explain how the human perceptual system interprets a scene and extracts data from it. Such proxies are a particularly useful reasoning tool for data visualizations, because understanding an individual’s—and a population’s—preferred proxies may suggest practical guidelines for how to optimize a visual representation to match these prox-

ies. This, in turn, would enable us to minimize the perceptual error arising from a specific visualization. Furthermore, proxies can also easily be operationalized as small programs (or “bots”) that model that proxy, which would allow us to estimate how effectively a given visualization should show a given pattern to a viewer.

9.1 Candidate Proxies

A visualization contains any number of visual features potentially available as a proxy for a given task, such as the lengths of the top most items of each set, or the perceived symmetry of each set. Different visual features might be better proxies than others for different visual comparisons. Here we explore which visual features appear to be most similar to participant performance (making the same decision), when used as a proxy for Maximum Mean or Maximum Range. We developed two broad categories of candidate features, informed by research in both visualization and perceptual psychology.

9.1.1 Global Features

Global-level features describe properties aggregated over a visual set of items, rather than comparing two focal items. Viewers can rapidly compute global statistics such as the mean of a collection of items [59, 60, 91, 92], though from present work it is unclear if this ability is mediated by a proxy. A list of hypothesized proxies of this type is shown in Figure 9.1. One high-precision proxy is that the lengths of bars in a set are genuinely averaged together and the chart with the largest ensem-








Possible proxies		Description	Visual cognition principle	
Global Proxies	Mean*		Extracts lengths of bars of each set, computes ensembles, chooses chart with longer ensemble.	People can extract the mean size of a set of items [25], though that mean is likely through some proxy.
	Centroid		Picks chart with largest centroid of the bar areas (along just relevant x axis).	Eye movements rapidly deploy to centroids of groups, but those centroids appear to be computed across the bounding hull of the objects, not their true center of gravity [17].
	Hull Area		Calculates convex hull of chart, picks chart with bigger hull area.	
	Hull Centroid		Calculates convex hulls, picks larger centroid (along relevant dimension only).	
	Trap Area		Draws trapezoids between each chart's top and bottom bars, picks bigger area.	A shape's external boundaries can be more visually salient than internal boundaries [6], which could produce overweighting of the first and last bars when judging the boundary contour.
	Trap Centroid		Draws trapezoids between each chart's top and bottom bars, picks trapezoid with larger centroid.	
	Symmetry Bias		Calculates skew (i.e. symmetry) of each set, chooses which chart is less skewed (i.e. more symmetric).	People are sensitive to symmetry [24] and are biased to select symmetric objects even when the task is not a symmetry judgment [15].

Figure 9.1: A set of “global-level” candidate perceptual proxies that might be used in visual comparison of means and ranges (and possibly other tasks).

ble length is chosen as the answer for the task. The mean length feature tests this genuine averaging. Viewers might also perceptually organize the bars into a coherent object, such that what they perceive is the convex hull of the bounded object that includes the heights of the bars and the white space between bars, and then compare the centroids or areas of these two hulls. These object boundary proxies might be subject to perceptual biases, such as overweighting outer edges in contour judgments [93]. Empirical research on human attention suggests that the allocation of attention throughout visual displays is preceded by the organization of the scene into objects and groups [94], and that the center-of-area of those objects can be rapidly computed [95]. The hull area and hull centroid proxies test whether this visual feature is consistent with participant responses and consistent with differences

in the data. Note that for superposed charts, the two hulls are overlapping, such that this particular visual feature may be harder for people to see because it involves filtering using color rather than space (as with the stacked, mirrored, and vertical arrangements). Finally, people are highly sensitive to symmetry in displays [20] and are biased to select symmetric over asymmetric information [96]. One possible heuristic is that people use symmetry as a proxy for range, such that any chart that is less symmetric is selected as the one having the bigger range.

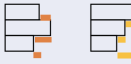





Possible proxies		Description	Visual cognition principle	
Focal Proxies	Range*		Extracts all pairwise deltas within a set, chooses set with the longest pairwise length difference.	Length differentiation has high acuity [16].
	Biggest Mover Pair (Abs)		Between charts, finds largest delta between item pairs (a1-b1, a2-b2...), picks chart with largest positive delta.	
	Biggest Mover Pair (Rel)		Same as Biggest Mover Pairwise, but scaled relative to the smaller item within the pair.	
	Biggest First Item		Compare top items, picks chart with larger top item.	When faced with multiple objects, people are biased to attend to the top object [26].
	Biggest Middle Item		Compare middle items, picks chart with larger middle item.	People might select a group's center item [17] and then select the larger of the two [16].
	Slope Min to Max		Finds each chart's min/max, computes slope between them. Picks the chart with the least-vertical slope.	People might select min/max outliers [13], then calculate offset [16].

Figure 9.2: A set of “focal” candidate perceptual proxies that might be used in visual comparison of means and ranges (and possibly other tasks).

9.1.2 Focal Features

Focal features describe pairwise differences between two items. People can discriminate small differences in line segment lengths [97]. Chart viewers might be sensitive to the deltas, either between charts (Biggest Mover Pair) or within a

chart (Neighbor Delta). In addition, focal attention can be biased to attend to the topmost item in a collection [62], so one possible proxy is that people compare only the lengths of the topmost items of the two sets (Biggest First Item). A list of hypothesized proxies of this type is shown in Figure 9.2.

9.2 Testing Proxies with Retrospective Analysis

To evaluate these proxies, we simulated what would happen if each proxy was tested on every data series combination that each observer actually saw in the Experiments 3 (Maximum Mean) and 4 (Maximum Range). Each proxy was used to make a decision about a visual comparison (e.g., Hull Area generated a convex hull around each of the two charts, calculated their areas, and evaluated the pixel difference in their areas), and provided an “answer” to the task (i.e., larger area is used as a proxy for mean or for range).

Note that this procedure necessarily shows the proxies different stimuli depending on arrangement: because the stimuli have been titrated to respond to viewer accuracy, the charts “shown” for stacked stimuli will have different properties than the charts “shown” for superposed stimuli. Because the data in the charts “shown” to the proxies is arrangement-specific, proxies were implemented to be arrangement-invariant. The proxies were calculated using raw data values, the length of each mark, and the relative location of each mark (e.g., the first datum in a chart was at the “top” location), not as visual features extracted from an image-based representation. Future work should also test proxy performance using

image-based implementations.

9.2.1 Implementation

We implemented these global and focal perceptual proxies for all charts.

We computed two outputs for each of these proxies: which chart would the proxy have chosen, and was this choice correct? Some visual features may be salient [98] to human observers, but not useful for an analytic task (uncorrelated with the answer). For example, the delta between adjacent bars (i.e., the amount of overhang) might be a salient and useful indicator for an analytic task involving comparing items, but if the viewer’s goal is to compare means, relying on this feature should impair task performance.

Although we excluded some participants from Experiments 3 and 4 for low accuracy, we included their data in the simulation to allow for the future possibility of testing whether their poorer task performance is consistent with using different perceptual proxies than other viewers with higher-precision visual comparison.

Proxies were implemented with Node.js, using D3 geometric libraries (though pseudocode below refers to a contrived “**geom**” library for generalization). For comparison to human decisions, the script was given as input the full list of trial data, containing, for each trial, the data points, the correct answer, and the answer chosen by the participant.

Basic statistical operations (mean, range, biggest mover, skew, biggest first item, biggest middle item, neighbor delta) are performed directly on the data, while

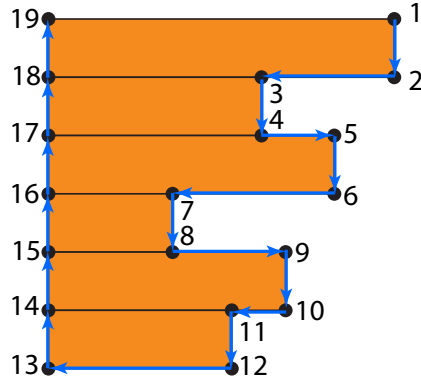


Figure 9.3: For geometric computations, points defining the boundary of the chart are enumerated in a clockwise manner. Note that the small white space between the bars when they are shown to human participants is not considered for these computations. Trapezoidal operations are performed using the points of only the first and last bars (1, 2, 11, 12, 13, and 19 in this example).

geometric operations (centroid, hull centroid, hull area, trapezoid centroid, trapezoid area) are performed on sets of points defining bar charts representing the data (Fig. 9.3, Alg. 5).

Algorithm 6 depicts the decision process for the hull centroid proxy. Others follow a similar paradigm; source code is available at <https://osf.io/uenzd/>.

Files that contain trial-by-trial data for properties of the stimuli, human responses, the pixel information used by each perceptual proxy to inform a heuristic about a chart decision, and each proxy's decision, for all combinations of arrangement and task, are also posted at <https://osf.io/uenzd/>.

Algorithm 5 Point generation

```
1: procedure POINTS( $d$ ) ▷  $d := \text{dataset}$ 
2:    $p \leftarrow []$ 
3:    $n \leftarrow \text{length } d$ 
4:   for  $i$  in 0 to  $n - 1$  do
5:     push  $p, [d_i, i/n]$ 
6:     push  $p, [d_i, (i + 1)/n]$ 
7:   for  $i$  in  $n$  to 1 do
8:     push  $d_{i-1}, [0, i/n]$ 
9:   push  $d, [0, 0]$ 
10:  return  $p$ 
```

Algorithm 6 Hull Centroid Proxy

```
1: procedure PROXYHULLCENTROID( $d_a, d_b$ ) ▷  $d_a, d_b := \text{datasets}$ 
2:    $p_a \leftarrow \text{Points}(d_a)$ 
3:    $p_b \leftarrow \text{Points}(d_b)$ 
4:    $c_a \leftarrow \text{geom.centroid}(\text{geom.convexHull}(p_a))$ 
5:    $c_b \leftarrow \text{geom.centroid}(\text{geom.convexHull}(p_b))$ 
6:   if  $c_a > c_b$  then
7:     return “a”
8:   else
9:     return “b”
```

9.2.2 Results

The goal of this proxy approach is to evaluate which visual features are consistent with human performance, and which are actually useful for the task. As such we evaluate the “decisions” of each proxy against two baselines. On what proportion of trials did the proxy agree with the participant’s response? And on what proportion of trials did the proxy agree with the true answer of the stimulus? We treat all of the following results as initial speculations, and make no claims of their statistical reliability. These values are depicted in Fig. 9.5. A visual feature can be considered useful if a decision using the differences in that visual feature is consistent with the task-dependent differences in the data. The dots in Fig. 9.5 to the right of 50% show proxies that give above-chance performance at the task. We highlight a few patterns.

First, the most useful proxies, in terms of finding the correct answer, depend on comparison task. For the Maximum Mean task, visual features of the Mean lengths (global), Bar Centroids (global), and Biggest Mover Pair (focal) were the most predictive of the difference in the means. It was unexpected that the Biggest Mover Pair, which computes pairwise differences between chart items, predicted the difference of means at above-chance levels. It suggests that in the data, the largest between-item change (neighbor delta in superposed charts, motion in animated charts) was predictive of the chart means, more so than other global features. For Maximum Range, the Range proxy (which computed all pairwise distances between items) was most useful, closely followed by pairwise differences only between

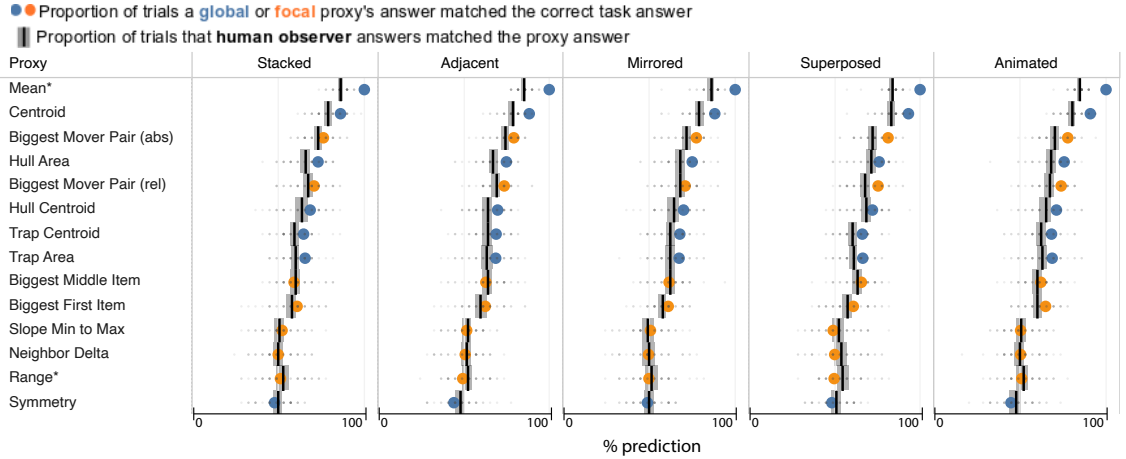


Figure 9.4: Results of the two analyses of visual proxy performance for the Maximum Mean task. The x-axis is the percentage of trials for which the visual proxy was predictive, for human behavior (vertical bars), and for true answer for the comparison (colored dots). The small dots show individual subjects, and light gray around the black lines shows 95% confidence interval. True answer dots are color-coded to show whether we informally coded them as a global proxy feature (blue) or focal proxy feature (orange). The true answer dots indicates that some features are more useful than others for a given visual comparison.

neighboring items (Neighbor Delta).

Second, people tend to make decisions consistent with using the most useful visual features: the bars that show agreement between proxy responses and human responses tend to follow the dots that shows the most task-relevant useful features in Figure 9.5.

Third, we note the absence of a symmetry bias. The Symmetry proxy, which uses stimulus symmetry as a proxy on which to make Maximum Mean and Maximum Range decisions, was predictive neither of actual differences in means or range, nor of human responses.

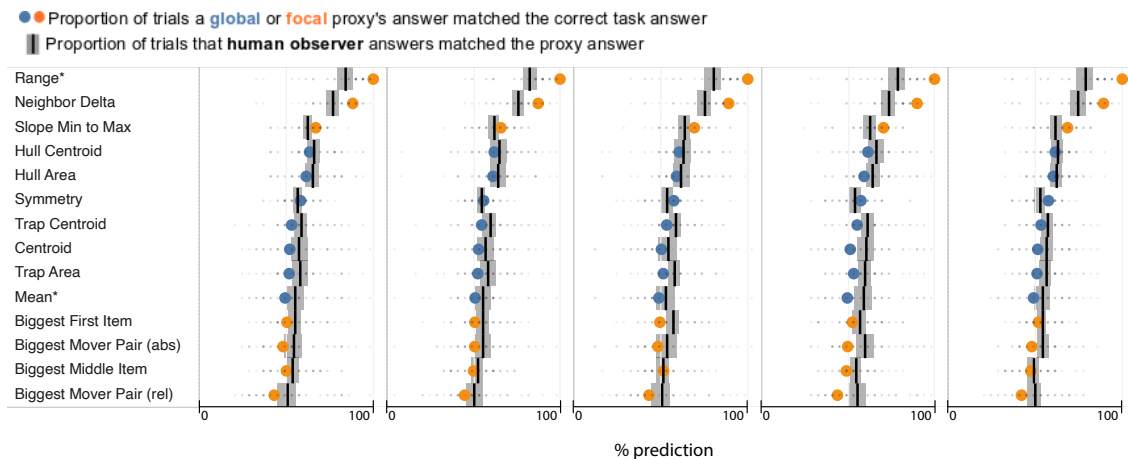


Figure 9.5: Results of the two analyses of visual proxy performance for the Maximum Range task. The x-axis is the percentage of trials for which the visual proxy was predictive, for human behavior (vertical bars), and for true answer for the comparison (colored dots). The small dots show individual subjects, and light gray around the black lines shows 95% confidence interval. True answer dots are color-coded to show whether we informally coded them as a global proxy feature (blue) or focal proxy feature (orange). The true answer dots indicates that some features are more useful than others for a given visual comparison.

Fourth, there is weak evidence of a bias for people to perform the Maximum Range task with the global proxies of Hull Centroid and/or Area Trapezoid Centroid, to a higher degree than is actually useful in the task: note where in Figure 9.5 the human behavior bars are to the right of the proxy dots.

We speculate that these findings are broadly consistent with the idea that global visual features are useful for set-level visual comparisons, and local visual features are useful for item-level visual comparisons. Maximum Mean and Maximum Range tasks benefit from the same chart arrangements, but use different emergent visual features in these chart arrangements for visual comparison. Visual comparison

is afforded by more than precision of marks and their arrangements. The “visual” component of visual comparison may rely on a flexible suite of visual proxies that viewers can rely on to accomplish a given task, depending on what visual features are present. The slight bias to erroneously use global features for the Maximum Range task raises the speculative possibility that, in some tasks and arrangements, viewers use global shape-based proxies even when these proxies are not useful.

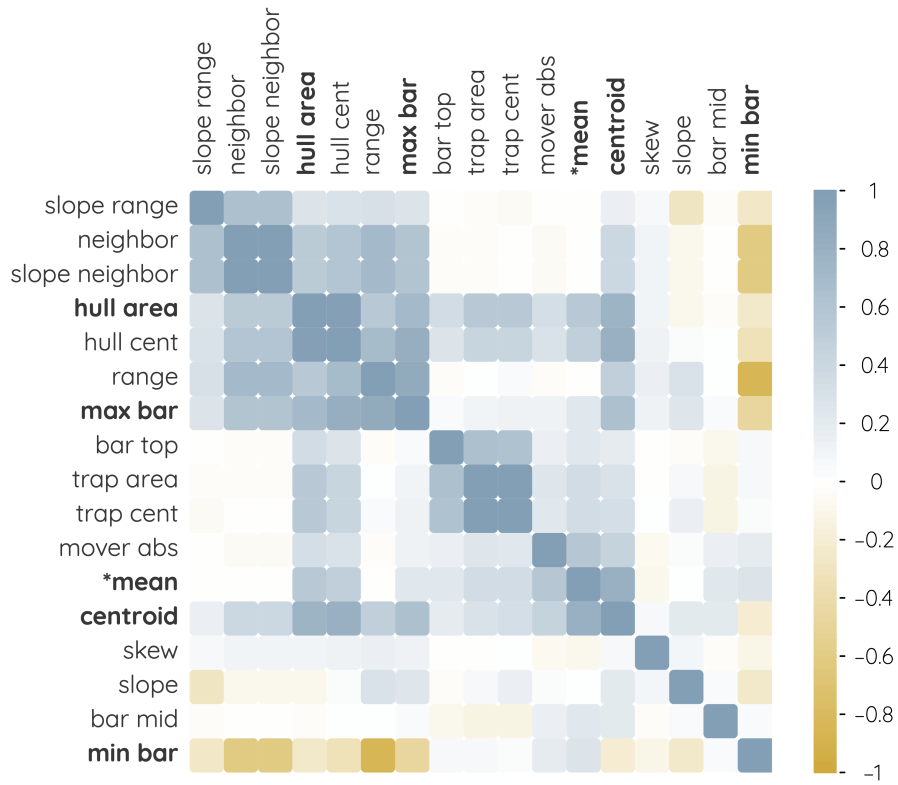


Figure 9.6: Proxy correlation for Maximum Mean data. In these data, many proxies (especially centroid) correlate well with the true answer (*mean), making it difficult for this retrospective analysis to distinguish use of these proxies from extraction of the true mean or some other proxies that correlate with the mean.

9.2.3 Limitations

The data analyzed here were generated to test comparative arrangements, rather than to tease apart proxies. As a consequence, many of the proxies were highly correlated in these data, both with each other and the true answer (Figs. 9.6 and 9.7). This makes it difficult to distinguish proxy effects.

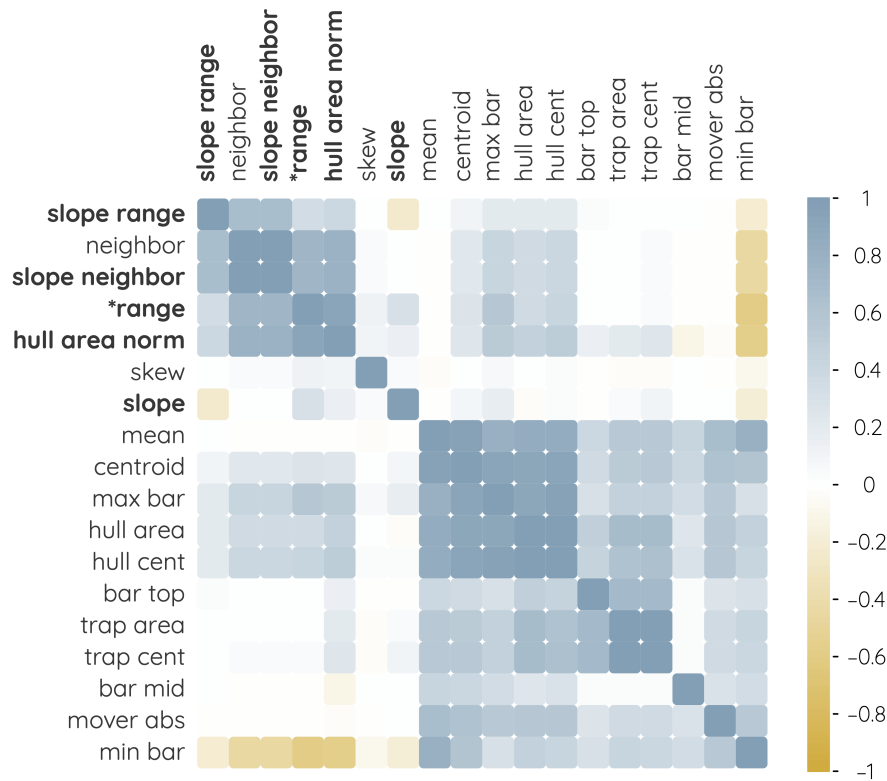


Figure 9.7: Proxy correlation for Maximum Range data. In these data, many proxies (especially hull area norm) correlate well with the true answer (*range), making it difficult for this retrospective analysis to distinguish use of these proxies from extraction of the true mean or some other proxies that correlate with the range.

Chapter 10: Revealing Proxies with Adversarial Examples

As discussed in Section 9.2.3, proxies tend to correlate with true summary values in data that is proxy-agnostic. This is not surprising, as charts would not be useful if the way we interpreted them had no bearing on their actual data. This poses a conundrum, however: if proxies always correlate with the true values, how can we determine, experimentally whether someone is using a proxy or computing the true value?

As a remedy, we propose searching for charts that are *adversarial*, which we define here as having a perceived summary statistic (i.e. proxy value) that deviates from the true value. Since the proxy values in these adversarial charts would not correlate with the true value, a preference for a given proxy would be a more robust indication that the proxy is used than in our retrospective analyses of random data.

10.1 Two Approaches: Testing vs. Learning

We will approach the problem of revealing perceptual proxies with adversarial examples in two complementary ways. (Fig. 10.1):

- **Theory-driven**, or “testing,” where we draw on the literature in vision science and visualization on perceptual proxies to generate “adversarial” datasets that

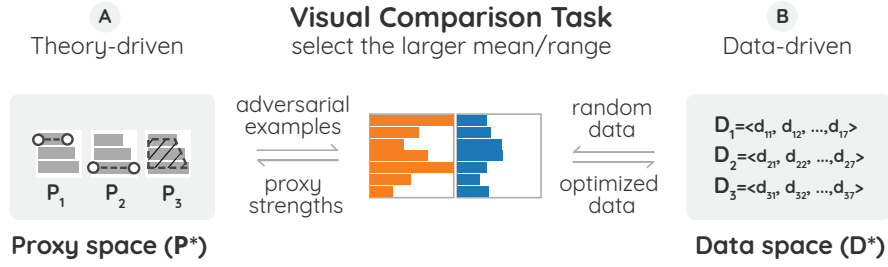


Figure 10.1: **Conceptual diagram of two adversarial approaches.** The theory-driven approach operates in proxy space, while the data-driven approach operates in data space.

optimize individual proxies to deceive a participant into selecting an incorrect choice (Experiment 5); and

- **Data-driven**, or “learning,” where we simply start from a set of randomly generated data series—with no preconceived notion of how they should be generated—and let participant choice for successive lineups between series guide a black-box optimization to find increasingly more deceptive data (Experiment 6).

10.2 Common Methods For Adversarial Experiments

A key feature of Experiment 6 (the “data-driven”, or learning, approach) is that all the charts needs to have the same true summary statistic. We suspected that participants may notice how similar they seem and resort to random guessing. It was thus important to run both experiments in parallel and in the same sessions, i.e., with the same participants (Fig. 10.2). All of the trials belonging to specific blocks—i.e., different proxies for Experiment 5 and different datasets being optimized for Experiment 6—were interleaved randomly. This way, participants would not know they were occasionally shown charts with the same mean or range and, in theory, continue to try their best. We only blocked the combined study on task type (see below), as each specific task requires specialized training.

Here we will discuss experimental aspect common to Experiments 5 and 6, including the two tasks (*MaxMean* and *MaxRange*), visual representation, apparatus, and procedure.

10.2.1 Visual Representation

We used a simple horizontal bar chart where each bar had a uniform color and thickness (Fig. 1.2). The visual stimulus involved showing these bar charts in a lineup consisting of two charts arranged side by side (i.e. the “adjacent” arrangement from Experiments 1-4). We used two diverging colors—orange (■ #ff7f0e) and blue (■ #1f77b4), respectively—for the two charts.

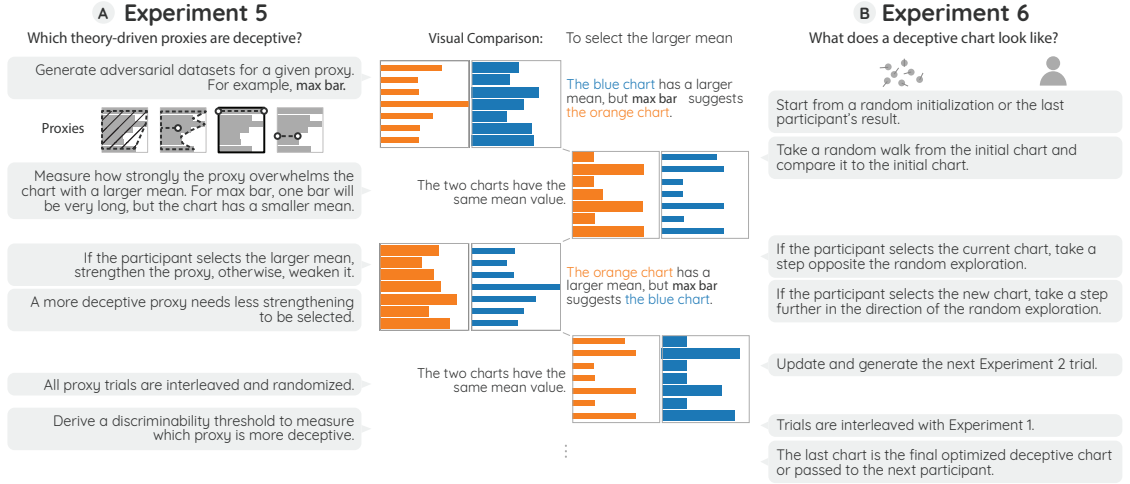


Figure 10.2: Interleaving of the two approaches. The two experiments are run in parallel with the same subjects, with trials from both interleaved as illustrated here. **A** In the “theory-driven” approach, we optimize charts to manipulate conjectured perceptual proxies, and test how powerfully they alter judgments. **B** In the “data-driven” approach, we seek to discover deceptive charts *de novo*, using human judgments as an objective function. The examples above present four real trials from the combined experiment. All annotations on bar charts are for illustrating purposes only.

10.2.2 Tasks

We test proxies using the same two tasks used in Experiments 3 and 4:

- **MaxMean**: Determine the chart that has the larger mean value across all of its components.
- **MaxRange**: Determine the chart that has the larger range from its shortest to its longest components.

While many tasks could be investigated, these are among the most basic of the summary statistics, yet are distinct from each other in that **MaxMean** is a “global”

task, since it requires computation across bars, while *MaxRange* is a “focal” task, since it requires the extraction of specific bars to compare, i.e., the min and max (see Chapter 9). Further, both these tasks are convenient because they can be computed on individual data series, unlike Maximum Delta (Experiment 1), which has dependencies across the two data series, and Correlation (Experiment 2), which requires two pairs of data series to create a forced-choice discrimination task. Note also that we use the “adjacent” arrangement even though these tasks were best supported by the “stacked” arrangement in Experiments 3 & 4 (see §7.3 and §8.3). In pilot experiments we found that stacked arrangements supported the tasks so well that it was difficult to discern any differences in performance, even with the maximum differences mathematically possible between the proxies and the true values. Thus, in this case, making the task harder (by using a sub-optimal arrangement) allowed us to push the visual system closer to its limits.

10.2.3 Procedure

After consenting, participants were shown a sequence of instructional screens followed by a set of practice trials. Practice trials gave feedback on whether or not the participant’s answer was correct; this was not the case for the timed trials. Participants were required to score three correct answers in a row to proceed past the practice phase. The purpose was to ensure that participants had correctly understood the task at hand.

As in Experiments 1–4, each individual trial started with a short countdown; then the platform showed the lineup of two data series visualized as bar charts in a side-by-side arrangement (horizontal juxtaposition) as impressions for a short time period. Based on extensive piloting, we chose 1000ms impressions for *MaxMean* and 1500ms for *MaxRange*. After the impression time ended, the lineup was replaced by two colored (orange and blue) buttons to represent the bar charts had been shown. Answering the trial meant clicking on the button representing the bar chart that the participant had perceived as having the larger mean or range. Participants assigned to each task typically spent between 8 and 27 minutes to complete all the sessions ($\mu = 15.24$, $\sigma = 4.75$).

10.2.4 Participants

For each of the two tasks, we recruited 65 participants for the combined study from Amazon Mechanical Turk (MTurk). The *MaxMean* task had 22 female, 42 male, and 1 unspecified, and the *MaxRange* task had 31 female and 34 male.

10.2.5 Apparatus

All experiments were distributed through the participant’s web browser. Because of our crowdsourced setting, we were unable to control the specific computer equipment that the participants used. We required a screen resolution of at least 1280×800 pixels. During the experiment, we placed the participant’s device in full screen mode to maximize the visibility. The testing software was implemented in

JavaScript and D3.js [\[87\]](#) with a server-side Perl and CGI backend.

Chapter 11: Experiment 5: Testing Proxies with Adversarial Charts

Experiment 5 follows a theory-driven approach: we start with a set of plausible perceptual proxies, generate datasets optimizing for them, and then test these datasets in human judgments. This experiment has two goals: first, to find evidence that participants could be using perceptual proxies in visual comparison tasks; second, to understand how participants used different proxies differently.

To find evidence of proxy use without entanglement of the true value, we create “adversarial” visualizations. These are pairs of charts for which the proxy would suggest a different answer than the true value in a forced-choice trial. The difficulty is controlled parametrically by a combination of the ratio of the true value (e.g. how much bigger is the correct mean) and how adversarial the pair of charts is (e.g. how much bigger the convex hull area is in the *wrong* chart). Both the values are encapsulated by a *titer* value.

11.1 Selecting Specific Proxies

We aimed to identify a set of proxies in the perceptual space that are likely used by participants and could be manifested by us to generate adversarial trials. We used the below heuristics and followed an iterative process. We primarily considered the


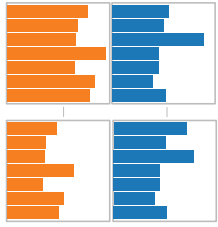
A <i>MaxMean</i>	
Confounding Proxy	Description
<i>ink area</i> 	<p>If the bar thickness is the same, the bar chart with a larger mean will always have more ink. In this example, the blue chart has a larger mean and more ink.</p>
	<p>We therefore vary the thickness. This example shows that the blue chart with skinnier bars could have less or the same amount of ink.</p>
B <i>MaxRange</i>	
Confounding Proxy	Description
<i>min bar and max bar</i> 	<p>In the orange chart, if we make max bar longer to be deceptive (a smaller range), min bar has to be longer, too. The blue chart will always have a shorter min bar.</p>
	<p>We span the range of the deceptive chart across min bar or max bar of the other chart and balance all the cases. In this example, the blue chart could have a longer min bar or a shorter max bar.</p>

Figure 11.1: The confounding proxies in the *MaxMean* and *MaxRange* tasks.

proxies that best aligned with participants’ judgments in our analyses in §9.2.2. We also considered a new proxy if it satisfies the above two constraints. Because of the high degree of correlation of many proposed proxies, we chose, from these, proxies that can be thought of as representatives of broader classes, based on qualitatively identifying clusters in correlation matrices (see Figs. 9.6 and 9.7). For example, the area of a chart’s convex hull is highly correlated with the horizontal position of that hull’s centroid, so we choose the former to represent the family of convex hull proxies, as it aligns slightly better with human choices in §9.2.2. Evidence for any proxy we have chosen thus would thus imply that either that proxy or a similar proxy is at play. As a result, we selected four proxies for the *MaxMean* task: *hull area*, *centroid*, *max bar*, and *min bar*; we also selected four other proxies for the *MaxRange*

task: *hull area norm*, *slope*, *slope range*, and *slope neighbor*. For each selected proxy, we show the description and an example in Figure 11.2 (*MaxMean*) and Figure 11.3 (*MaxRange*).



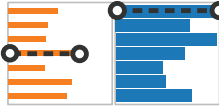

<i>MaxMean</i>		
Proxy		Description
<i>hull area</i>		The area of a convex hull around the bars, ignoring difference in bar thickness.
<i>centroid</i>		The centroid of the area occupied by the bars along just relevant x-axis
<i>max bar</i>		The length of the longest bar
<i>min bar</i>		The length of the shortest bar

Figure 11.2: **Perceptual proxies used for *MaxMean* adversarial experiments.**

All the example chart pairs have the same underlying datasets, and the blue chart on the right side has a **larger mean** (the correct answer). In trials, the position of the correct answer is randomized and balanced. Charts randomly have skinny bars to decouple amount of ink from the mean (see Section 11.2).

11.2 Eliminating Confounding Proxies

Besides the selected proxies, both tasks had other proxies directly related to the summary statistic itself. They could always indicate a correct answer (i.e., the larger mean or the larger range), and thus we attempted to eliminate their impact in our experiment.



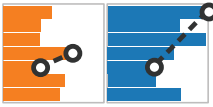
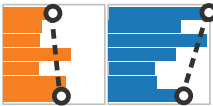
<i>MaxRange</i>		
Proxy		Description
<i>hull area norm</i>		The area of a convex hull around the bars, cropped to the shortest bar
<i>slope neighbor</i>		The largest slope from the tip of one bar to the tip of an adjacent bar
<i>slope range</i>		The slope from the tip of the minimum bar to the tip of the maximum bar
<i>slope</i>		The slope of a regression line fit to all the bars

Figure 11.3: **Perceptual proxies used for *MaxRange* adversarial experiments.**

All the example chart pairs have the same underlying datasets, and the blue chart on the right side has a **larger range** (the correct answer). In trials, the position of the correct answer is randomized and balanced. Slopes are computed using bar length as “height” and distance between the centers of the bases of the bars as “width.”

For the *MaxMean* task, an *ink area* proxy—the total “amount of ink” [34] (i.e., the number of colored pixels on the screen)—could be used by participants to estimate mean when the number of bars is different [25]. If all the bars are of the same thickness, the *ink area* proxy reduces to the sum, and thus the arithmetic mean (see Table. 11.1a). We decoupled the *ink area* proxy from the mean by randomly choosing one of the two charts to have skinnier bars than the other. We chose a fixed skinniness such that the skinny-bar chart will always have the least amount of ink, even for a large difference in mean. The *ink area* value thus cannot be used to determine the correct answer.

Similarly, for the *MaxRange* task, *min bar* and *max bar* are closely related to the range (see Table. 11.1b). Therefore, the other chart will always have a shorter *min bar*. The feedback from the pilot studies also supported this speculation, as some participants reported choosing the chart with the shortest bar as their strategy. We therefore manipulated the range values such that the smaller range spans either the minimum or maximum of the larger range. In this way, *min bar* or *max bar* only corresponds to the larger range 50% of the time and therefore is no longer correlated with the correct answer.

For each of these confounding proxies, we randomized and balanced the four cases: if the proxy is deceiving or not and if the correct response is on left or right.

11.3 Hypotheses

With our goal of understanding proxies and participants’ usage of specific proxies, we framed two research hypotheses for Experiment 5:

- \mathcal{H}_1 Adversarially manipulating perceptual proxies will mislead participants to be *worse* at making a visual comparison.
- \mathcal{H}_2 Individuals will be affected by such manipulations differently.

11.4 Experimental Design

For our hypotheses, we performed within-subjects factorization for the two tasks and the corresponding four proxies. We recruited different participants for each task due to concerns about practice [99], fatigue [100], and carryover effects.

Each participant finished all four proxy conditions and a control condition where no specific proxy was manipulated. We designed this control condition to replicate the results from Experiments 5 and 6 and also to provide a baseline for comparison. Each condition consisted of 20 trials. In each trial, we collected the participant’s response, the proxy manipulated, the two datasets presented, and the experiment parameters. The remaining details of experimental materials, framework, recruitment, procedure, and data collection were described above in Sections 11.5 and 10.2.

11.5 Generating Adversarial Charts with Simulated Annealing

We generate adversarial pairs of charts for a given proxy using simulated annealing [101], drawing inspiration from Matejka and Fitzmaurice [90]. Our objective is a pair of datasets with specified ratios for a proxy and a summary statistic. Deviation from this objective is formalized in a cost function as the sum of squared differences between the ratios in the objective and those of the dataset being considered, as in Equation 11.1. Here $\mathbf{x}^{(i)}$ is a vector of the bar lengths for the chart $i = 1, 2$. The functions $\mu(\mathbf{x})$ and $p(\mathbf{x})$ represent the true statistic (e.g. mean) and proxy function (e.g. convex hull area), respectively. In addition to bar lengths $\mathbf{x}^{(n)}$, $n = 1, 2$, the function takes target values for the statistic (μ'_i) and proxy (p'_i) in each of the two charts ($i = 1, 2$).

$$J(\mathbf{x}^{(n)}, \mu'_n, p'_n \mid_{n=1}^2) = \sum_{i=1}^2 (\mu(\mathbf{x}^{(i)}) - \mu'_i)^2 + (p(\mathbf{x}^{(i)}) - p'_i)^2 \quad (11.1)$$

11.6 Measurement

To manifest specific proxies and quantify their effects on the *MaxMean* and *MaxRange* tasks, we followed the methodology of Experiments 1–4. Following these, the *titer* value for a pair of bar charts (left and right) is defined as follows:

$$titer = \frac{\max(S_{\text{left}}, S_{\text{right}})}{\min(S_{\text{left}}, S_{\text{right}})} - 1, \quad S \in \{f_{\text{mean}}, f_{\text{range}}\} \quad (11.2)$$

where S is a summary statistic for the dataset, and it could be arithmetic mean (f_{mean}) or range (f_{range}). The *titer* value normalizes the difference of a summary statistic for the two side-by-side bar charts and scales task difficulty in different trials. For example, if a *titer* value is 0.1 in a *MaxMean* trial, one of the bar charts has a mean value 10% larger than the other one in homogeneous coordinates. In practice, a *titer* value of 0.5 is considered very large for participants to tell the larger mean or range. Examples of various *titer* values can be seen in Figures 11.4 and 11.5.

If participants need a large *titer* to correctly discriminate the summary statistic between the two bar charts (e.g., they need more differences in mean to select the larger mean), they are more likely to be deceived by the adversarial examples towards an incorrect answer, and therefore they likely use those proxies. Alternatively, if participants successfully select the correct answer with a small *titer*, they may not be deceived by our manipulation of proxies.

We seek a *titer threshold* to summarize all the trials in an experimental condition and to describe participants' performance for that condition. The *titer threshold*

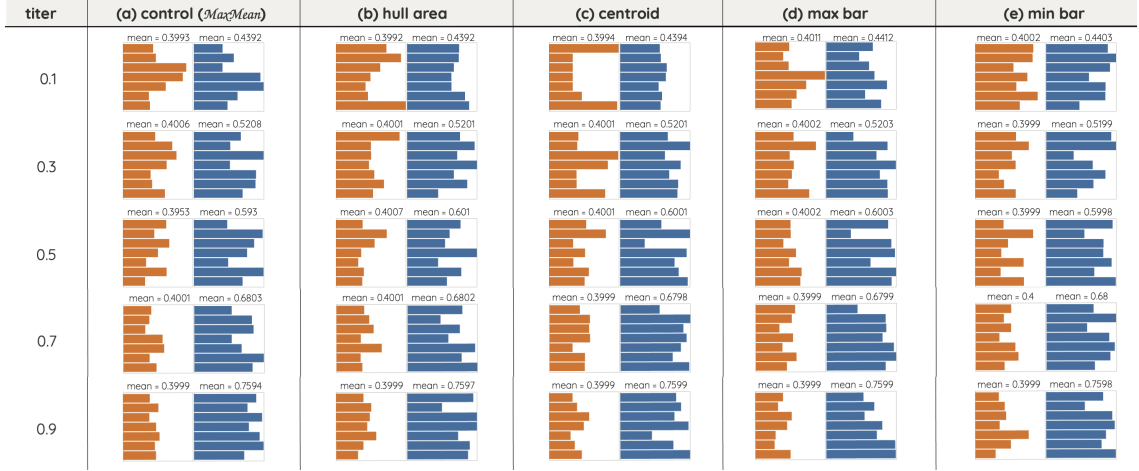


Figure 11.4: **Example titers for $MaxMean$.** We show examples of bar charts for each condition, including the control condition. Note that, for each trial, charts are parametrically generated and will not be the same as these. In each proxy condition, we optimize that particular proxy and make it more deceptive. The bar charts on the left side have roughly the same mean value around 0.4. The bar charts on the right side always have a mean value higher than 0.4, which are the correct answers. Note that to facilitate comparison, we use the same thickness for all the bars.

describes when participants could just discriminate the difference ratio of a summary statistic. This threshold concept is similar to the concept of *discrimination threshold*, like a *just noticeable difference* (JND) [102], but we use a difference ratio rather than absolute difference to normalize the stimuli. To measure a titer threshold, we started with titers of 0.25 and 0.40 for the $MaxMean$ and $MaxRange$ tasks, respectively, and approached the threshold using a staircase method [103]. The staircase method increased the *titer* value for an erroneous response (making the next trial easier) and decreased for a correct one (making the next trial more difficult) with two stages: in the first four trials, the increment and the decrement were both 0.03

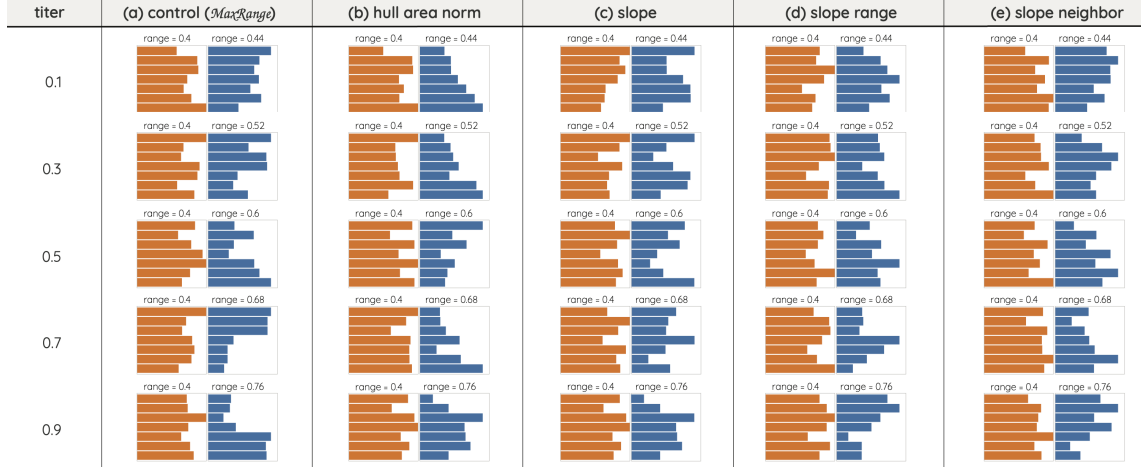


Figure 11.5: **Example titers for *MaxRange*.** We show examples of bar charts for each condition, including the control condition. Note that, for each trial, charts are parametrically generated and will not be the same as these. In each proxy condition, we optimize that particular proxy and make it more deceptive. The bar charts on the left side have the same range 0.4. The bar charts on the right side always have a range value larger than 0.4, which are the correct answers.

for *MaxMean* and 0.06 for *MaxRange*; in the rest of the trials, the decrement was 0.01 for *MaxMean* and 0.02 for *MaxRange*. These mechanisms ensure that we efficiently present stimuli to participants and conceptually align with measuring 75% JND; that is, the minimum difference (ratio) could be reliably discriminated 75% of the time [74, 102].

11.7 Prerequisites for Analysis

Data We planned to include all the participants and analyze all their data. We made only one exception where we excluded one participant from the *MaxMean* task

due to an assignment error. As such, for the *MaxMean* task, we based our analysis upon 6,400 trials = 20 trials per condition \times (4 + 1) conditions \times 64 participants; and for the *MaxRange* task, we based our analysis on 6,500 trials = 20 trials per condition \times (4 + 1) conditions \times 65 participants.

Replication Our two control conditions were similar to the “adjacent” conditions in Jardine et al. [26], and the number of participants (65) was also similar to theirs (50). To compare our results with theirs, we followed the same analysis method to calculate the average of titer values in the last ten trials and 95% confidence intervals from a Student’s *t*-distribution. As a result, we had 0.19 [0.17,0.21] for *MaxMean* and 0.46 [0.41,0.51] for *MaxRange*, compared to 0.21 [0.19,0.24] and 0.32 [0.30,0.33] from Jardine et al. While our *MaxMean* results are similar to Jardine et al.’s, our *MaxRange* task appeared to be more difficult. This may be because we mixed the control condition with other adversarial trials and trials from Experiment 2.

Bayesian estimation For our own analyses, we followed a Bayesian estimation approach [104, 105]. We used weakly informative priors to incorporate constraints of the experimental design and to roughly capture theoretically possible values within two standard deviations. We used the R packages *brms* [106], *ggdist* [107], *tidybayes* [108], *rstan* [109], and *tidyverse* [110] for computing and presenting the results.

11.8 Analysis

Our analysis had two steps. First, we used separate Bayesian logistic regressions directly on participants' responses to estimate each participant's titer threshold for each proxy. From these models, we also derived the measurement error of participants' thresholds. Second, we used the titer thresholds and measurement errors in a robust Bayesian mixed-effects linear regression to estimate the effects of each proxy on participants' perception.

This two-step analysis protocol aligns with a common approach to aggregating repeated trials when analysing JNDs (e.g., [26, 74]), but also incorporates measurement error from the first models into the second to reduce variance. From the results of the second model, we compare different perceptual proxies (\mathcal{H}_1) and infer their various effects on different individuals (\mathcal{H}_2).

11.8.1 Step 1: Deriving Thresholds and Measurement Error

We illustrate how we derive titer thresholds and the associated measurement error in Fig. 11.6.

Logistic regression For each proxy \times participant, we built a Bayesian logistic regression model for that participant's 20 dichotomous responses on that proxy (1 if the participant correctly selected the chart with the larger mean/range, 0 otherwise) (Fig. 11.6a). The resulting logistic curves describe the relationship between titer values and the probability of a participant making a correct response (between 1

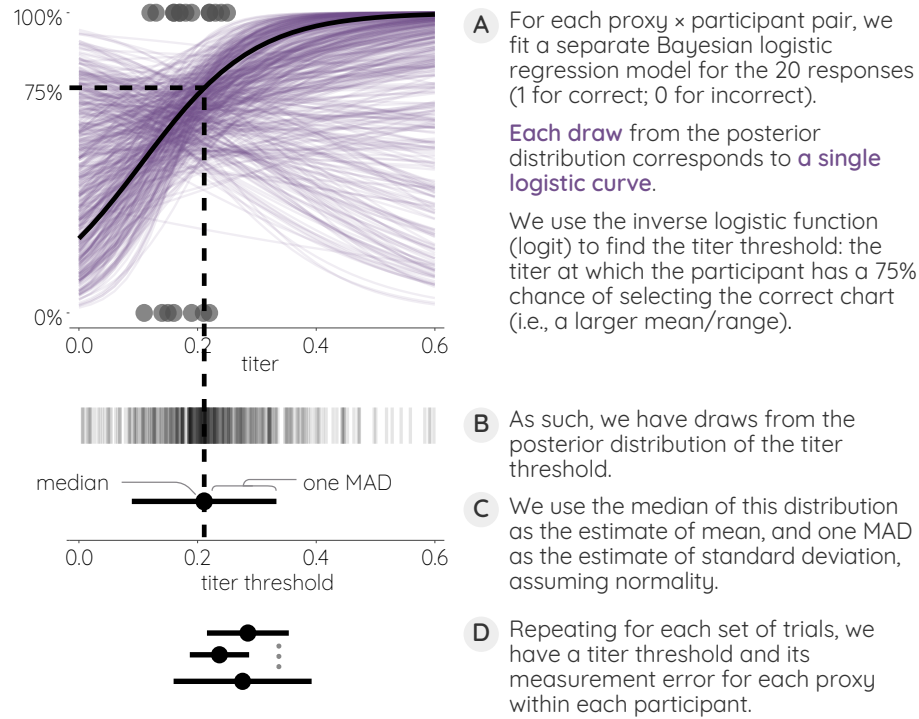


Figure 11.6: **Deriving titer thresholds and measurement error.**

and 0). We used the inverse logistic function (logit) to calculate the corresponding titer value at which a participant has a 75% chance of getting the correct response; this value is the *titer threshold*. Similar approaches are common in psychophysics to calculate JNDs [102], and have recently been used in visualization [24, 111].

Measurement error Because we use two steps to our modeling (logistic regression to find titer thresholds followed by a linear model of thresholds), there is *measurement error* [112] associated with the titer thresholds that should be propagated from the first models to the second: the titer thresholds are uncertain, as they are estimated from data. In a Bayesian context, we can propagate this measurement error by replacing the point estimates of titer thresholds with probability distributions [113]. From the posterior distribution of each logistic regression model, we

use robust estimates of location and scale—median and median absolute deviation (MAD) [114]—to derive a titer threshold (μ_{ij}) and the associated measurement error (σ_{ij}) for each participant $i \times$ proxy j (Fig. 11.6b). Then, in the linear regression (described below), instead of a response variable consisting only of point estimates (i.e., just the estimated titer thresholds, μ_{ij}), our response variables are distributions: $\text{Normal}(\mu_{ij}, \sigma_{ij}^2)$. This is a straightforward approach to measurement error in a Bayesian context [113].

11.8.2 Step 2: Modeling Thresholds

Mixed-effects linear regression We used a robust Bayesian mixed-effects linear regression to model the titer thresholds. We used a Student’s t distribution instead of a Normal distribution as the likelihood to make the model more robust to outliers [115]. We followed a measurement error approach and specified our response variables as Normal distributions corresponding to titer threshold estimates and their measurement error (see Step 1 above). We specified *proxy* as a fixed effect, so that different proxies can have different titer thresholds on average. We then used a random intercept and random slopes for *proxy* dependent on *participant*. This allows each participant to have their own titer thresholds within each proxy in the model. In **brms**’s [106] extended Wilkinson-Rogers [116] notation, this model is:

$$titerThreshold|se(titerError) \sim proxy + (proxy|participant) \quad (11.3)$$

Where *titerThreshold* is the estimated titer threshold (μ_{ij} above), *titerError* is the measurement error in the titer threshold (σ_{ij} above), and *proxy* and *participant* are

categorical variables indicating the manipulated proxy and participant, respectively.

11.9 Results

We report medians, 50% and 95% quantile credible intervals (CIs; Bayesian analogs to confidence intervals) as estimates of mean effects, and present the medians of posterior predictive distributions to show individual differences, following the presenting style of Fernandes et al. [117] and Hullman et al. [118].

11.9.1 The Effects of Manipulating Perceptual Proxies

We report here the mean effects for each proxy and comparisons with the control condition (no proxy was manipulated). We found evidence to support \mathcal{H}_1 : participants are likely deceived by some of the manipulated proxies.

MaxMean (Fig. 11.7) The four proxies have posterior distributions surrounding and similar to the control condition. When looking at the posterior distributions of differences in titer threshold, weak evidence supports that manipulating *centroid* might lead to a *larger* average titer threshold, suggesting that an average participant might be *deceived* by the *centroid* proxy, and therefore might be using that proxy to estimate *MaxMean*. Manipulating *hull area*, *max bar*, or *min bar* is less likely to have a large effect on average, suggesting that an average participant is less likely to be deceived by those proxies.

MaxRange (Fig. 11.8) We did not find strong evidence of an effect of either *hull area norm* or *slope neighbor* on titer threshold. The *slope neighbor* proxy is most likely to lead to

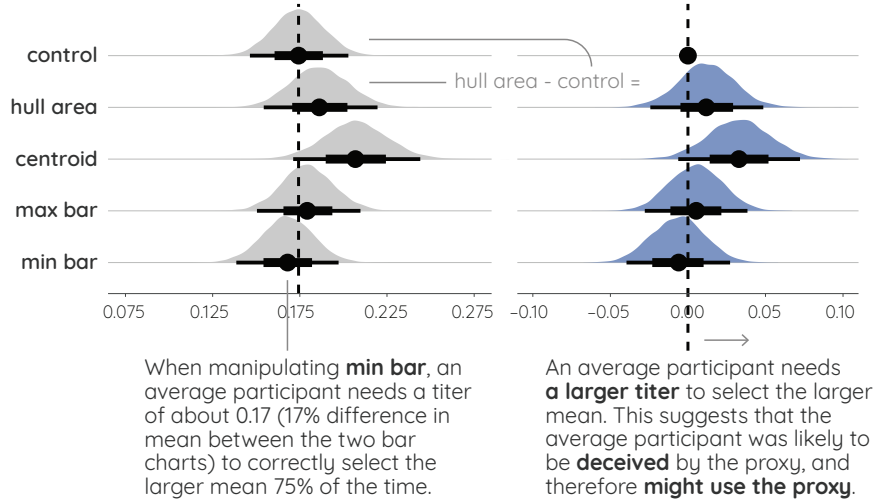




Figure 11.7: **The effects of manipulating perceptual proxies (\mathcal{H}_1) for $MaxMean$.**

We show posterior distributions (), 50% and 95% CIs () of expected titer thresholds (x-axes; the titer value at which 75% accuracy is expected), and a comparison with the control condition. Plots to the right (colored) show the same values as the left, but as offsets from the mean of the control conditions.

larger titer thresholds, but neither the chance of this nor the associated size of the effect are large. We found *slope* and *slope range* are likely to yield smaller titer thresholds, suggesting that an average participants is more likely to *select against* these two proxies. As we explain below, this may suggest the presence of some other proxies, negatively correlated with *slope range* and *slope neighbor* (proxy conflicts), which an average participant might be using.

11.9.2 Interpreting Participants Selecting Against a Proxy

We found that *slope range* and *slope* might lead to smaller titer thresholds on average than the control condition. When this happens, we say that participants are *selecting against* a proxy. Consider two bar charts, \mathcal{A} and \mathcal{B} (see Figure 11.9).

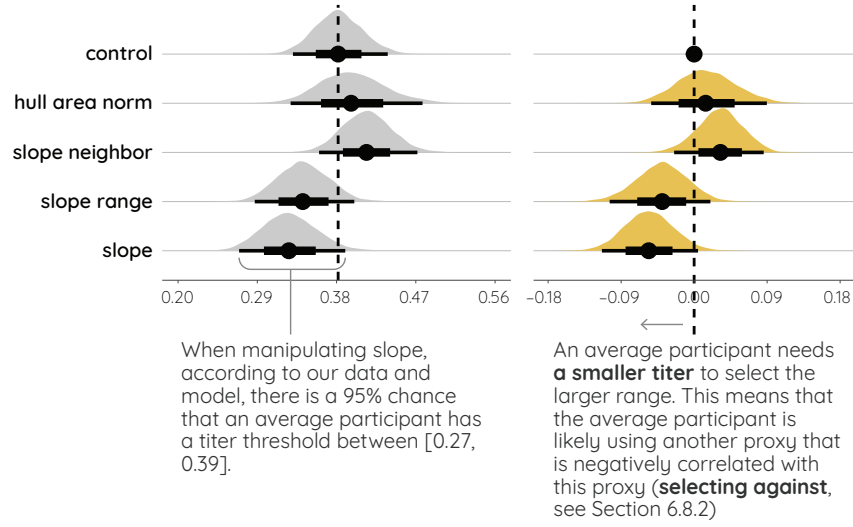


Figure 11.8: **The effects of manipulating perceptual proxies (\mathcal{H}_1) for *MaxRange*.**

We show posterior distributions (\triangle), 50% and 95% CIs (\bullet) of expected titer thresholds (x-axes; the titer value at which 75% accuracy is expected), and a comparison with the control condition. Plots to the right (colored) show the same values as the left, but as offsets from the mean of the control conditions.

\mathcal{A} has the larger *slope* (our manipulated proxy) and the smaller range of the two; \mathcal{B} has the smaller *slope* but the larger range. Say participants do not use *slope*, but do use some other proxy \mathcal{Y} that is negatively correlated with *slope* (e.g., *slope neighbor*), such that \mathcal{B} has the larger value of \mathcal{Y} . Now \mathcal{B} has both the larger value of \mathcal{Y} and the larger *slope*, so participants using proxy \mathcal{Y} will be more likely to correctly pick \mathcal{B} at a smaller titer, leading *slope* to have a smaller titer threshold than the control. Thus, the smaller titer thresholds of *slope range* and *slope* suggest there may be some other proxy (negatively correlated with *slope range* or *slope*) that participants were using.

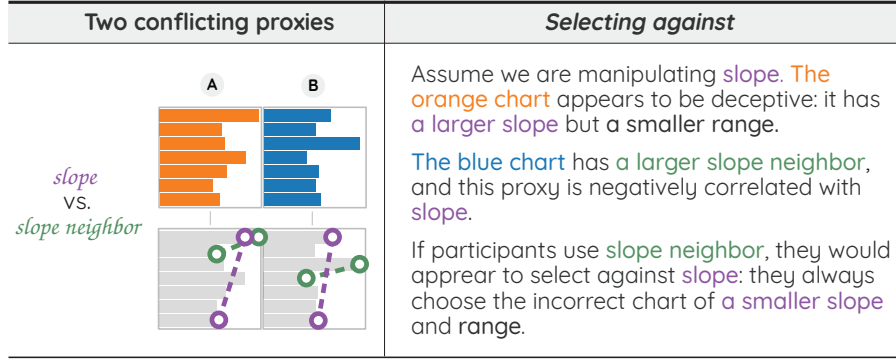


Figure 11.9: An example of participants selecting against a proxy.

11.9.3 Individual Differences

To investigate individual differences, we report each participant’s median of predicted expected titer threshold and a comparison to the control condition across different proxies in Figs. 11.10 and 11.11, assuming no measurement error. We found evidence supports that participants use proxies differently for our \mathcal{H}_2 .

MaxMean (Fig. 11.10) We found that on average, most participants are consistent with themselves across all conditions (③): participants who have larger titer thresholds than others in the control condition are more likely to have larger titer thresholds in other conditions and vice versa. This is reasonable: if participants are good at selecting the larger mean between the two charts, they could have been good at the task across different conditions, and thus result in smaller titer thresholds in all the conditions. A large portion of participants behave similarly (④), but a small portion of participants have larger titer thresholds than the others.

We found that most participants seem to be deceived by the adversarial trials, suggesting that they might use the manipulated proxies or other proxies positively

correlated with these. The exception is that in the *min bar* condition, participants seem to be consistently and slightly selecting against our manipulation, indicating that they might use other proxies negatively correlated with *min bar*. A handful of participants seem not to follow any manipulation (⑥); their titer thresholds are similar to those of the control condition. Different participants are likely to be deceived by different proxies to different extents (⑦). While the majority of participants seem to be deceived by *centroid* the most (⑧), *centroid* is also where participants' behavior deviate from each other the most. Last, participants are more similar across and within *min bar* and *max bar* conditions, meaning that in our procedure, if participants use the *max bar*, they are less likely to use *min bar*, consistent with our observations from Section 11.9.2.

MaxRange (Fig. 11.11) We found that most participants appear to be self-consistent across all the proxy conditions (①), but less consistent than those participants in the *MaxMean* task (they were different participants). Participants who have larger titer thresholds than others in the control condition are more likely to have larger titer thresholds in other conditions (②) and vice versa (③). These two groups appear to have similar numbers of participants, and there are other participants who behave differently across different conditions (④).

We found evidence supports that participants might use different proxies differently across different conditions. Participants are most similar to each other in *slope neighbor* (⑤); but they are least similar in *full area norm*. Some participants could be deceived by the manipulated proxy, while some are selecting against the proxy, and others are likely not to follow the manipulation; most participants are

likely selecting against both *slope* and *slope range*. Different participants may ignore a manipulated proxy, be deceived by a second one, but select against another one ((6)-(8)).

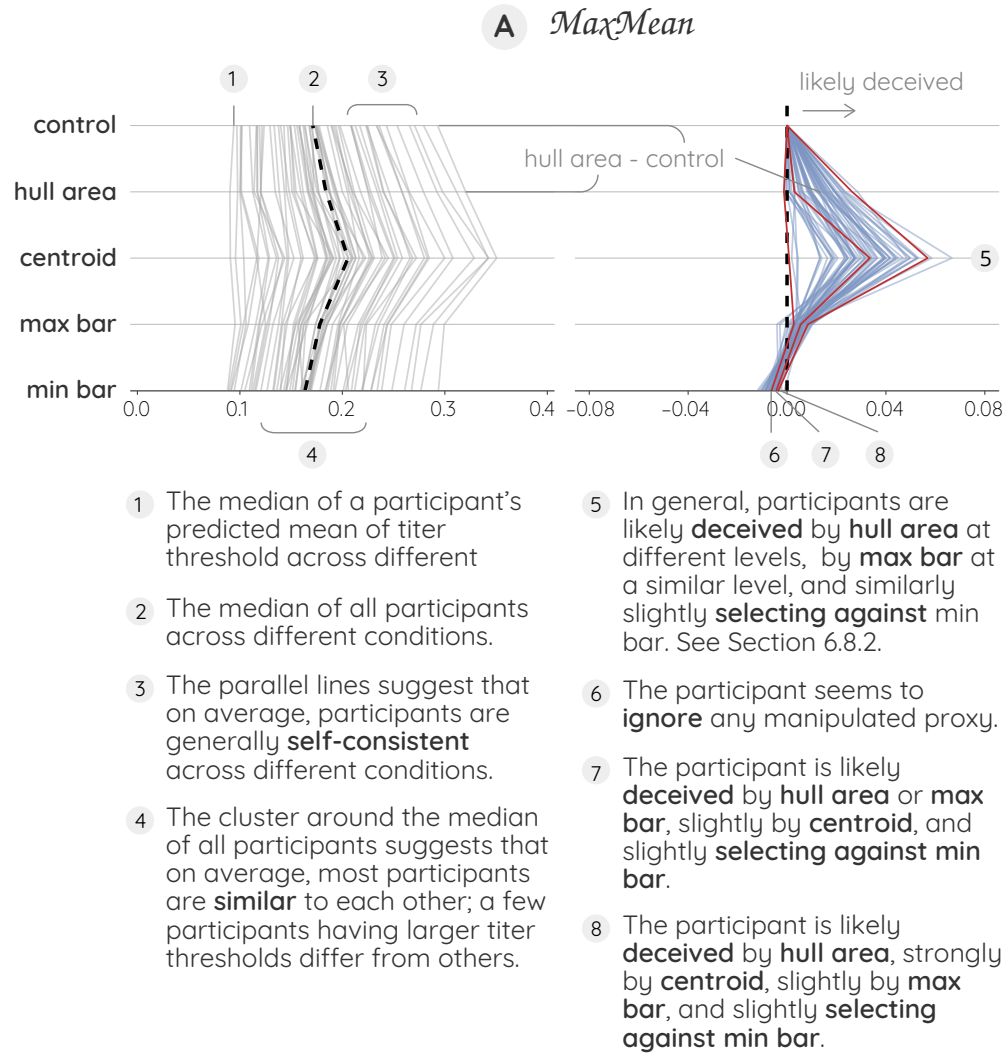


Figure 11.10: **The individual differences in different proxies conditions (\mathcal{H}_2).**

We show posterior predicted median of titer thresholds and a comparison with the control condition for each participant (\gtrsim).

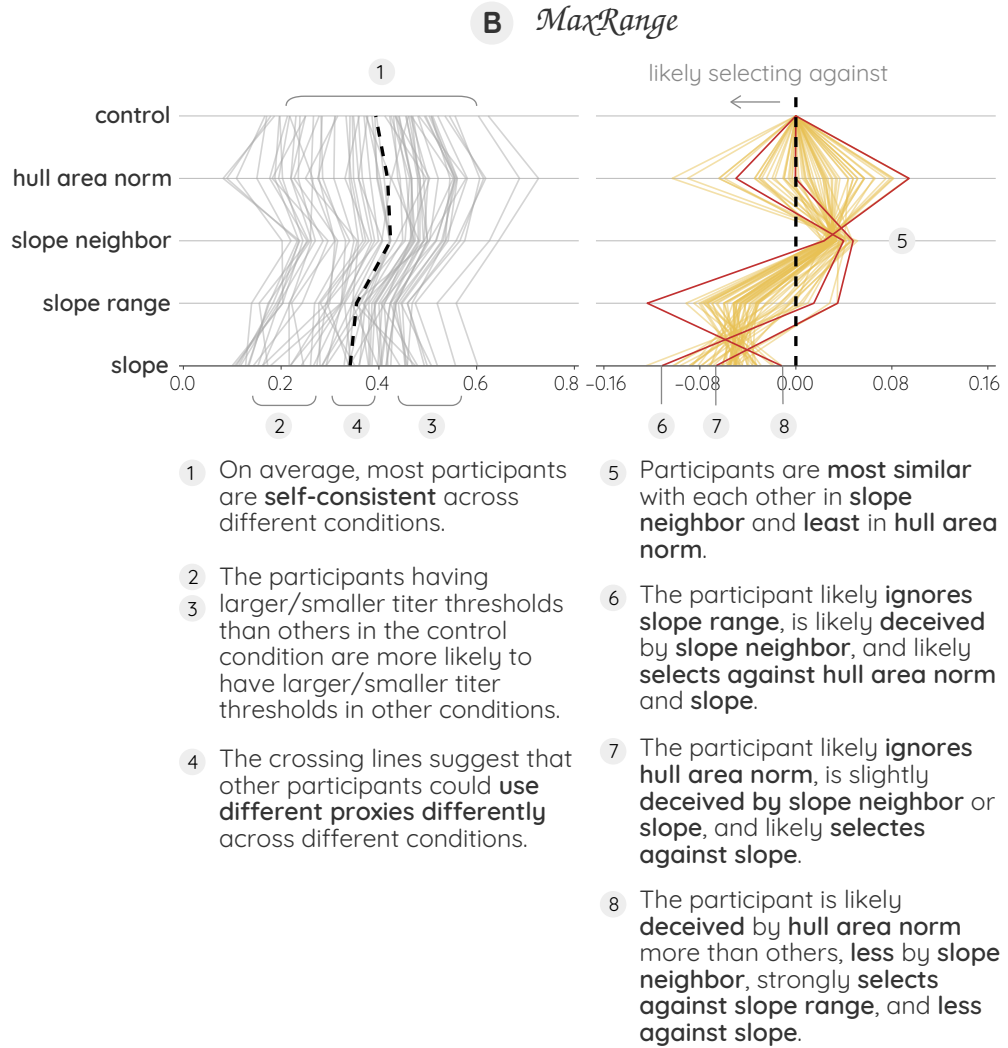


Figure 11.11: **The individual differences in different proxies conditions for *MaxRange* (\mathcal{H}_2).** We show posterior predicted median of titer thresholds and a comparison with the control condition for each participant (\gtrsim).

Chapter 12: Experiment 6: Learning Adversarial Charts Interactively

In Experiment 5, we started from the assumption that specific proxies may be at play and attempted to probe their effects. While this allowed direct testing of these hypotheses, it also, necessarily, restricted the types of data participants saw. Further, it could not provide us with any information about the effects of other proxies that were not tested, either because we did not deem them plausible or we did not conceive of them when generating our initial list. In a second adversarial experiment, we thus approach the question from the opposite direction, asking instead: what kinds of datasets *appear* to have a larger mean or range? More specifically, we consider the human perception of the summary statistic to be a black-box function [119] that we are seeking to optimize. In other words, beginning from only random data, we seek to use human judgments to guide the transformation of those data into charts that are deceptive with regard to a summary statistic. We will operate under the assumption that if a participant is asked to pick a chart with a higher summary statistic (i.e. mean or range), but those charts actually have the *same* summary statistic (unknownst to them), then they will pick the chart that has the higher *perceived* summary statistic. We further assume that such perception

is independent of (though likely correlated with) the actual summary statistic, due to the use of some perceptual proxy.

We frame two hypotheses for Experiment 6:

- \mathcal{H}_1 Optimized charts will display identifiable characteristics corresponding to the proposed proxies.
- \mathcal{H}_2 Optimized charts will be adversarial, appearing to have larger summary statistics versus random charts with the same statistics.

12.1 Optimization Method

Since we do not have access to the hypothetical “function” that describes human perception of summary statistics (let alone derivatives), we implement Dueling Bandit Gradient Descent (DBGD) [120], which stochastically estimates the gradient descent process using only pairwise rankings. We optimize in data, or “bar,” space, meaning we have 7 dimensions representing the lengths of each bar, in order from top to bottom. Our version of the algorithm is depicted in Algorithm 7. This method requires a projection function \mathcal{P} to map points from euclidean space to the feasible set for a given optimization problem. In this case, the feasible set is all charts with the same mean μ (for *MaxMean* task) or the same range $[min, max]$ (for the *MaxRange* task). Taking steps along random vectors, as required by the algorithm, typical moves the query point outside of the feasible set (i.e. it changes the mean or range of the chart), and the projection allows the algorithm to stay as close as possible to the chosen point while satisfying the constraints. The estima-

tion methods we use for these projections to a given mean (\mathcal{P}_m) or range (\mathcal{P}_r) are given in Algorithms 8 and 9, respectively. Note that the *data* space we explore lies between 0 and 1 in each dimension. However, since our charts have a minimum bar length of 0.25 times the chart width, data are mapped from $[0,1]$ to $[0.25,1]$ when converting to *bar* space.

Algorithm 7 The Dueling Bandit Gradient Descent algorithm. We deviate slightly from the algorithm as originally published in the case where the new point is not better than the current one: rather than doing nothing, we instead take a small step opposite the random exploration vector \mathbf{u} (line 15), both so that progress is made faster and so that participants do not see repeats of identical charts. This introduces a new parameter, η to denote the magnitude of this backward-step vector.

```

1:  $n := \text{dimensions}$  (7)

2:  $\mathbf{d}_0 := \text{initial point}, \in [0, 1]^n$ 

3:  $\delta := \text{exploration step size}$  (0.5)

4:  $\gamma := \text{forward exploitation step size}$  (0.9)

5:  $\eta := \text{backward exploitation step size}$  (0.1)

6:  $k := \text{iterations}$  (20)

7:  $\mathcal{P} := \text{project to feasible set}$ 

8: procedure DBGD( $\mathbf{d}_0, \delta, \gamma, \eta, k$ )

9:   for  $i = 1..k$  do

10:      $\mathbf{u} \leftarrow \text{uniformRandomUnitVector}()$ 

11:      $\mathbf{d}' \leftarrow \mathcal{P}(\mathbf{d}_{i-1} + \delta \mathbf{u})$ 

12:     if  $\mathbf{d}' \succ \mathbf{d}_{i-1}$  then

13:        $\mathbf{d}_i \leftarrow \mathcal{P}(\mathbf{d}_{i-1} + \gamma \mathbf{u})$ 

14:     else

15:        $\mathbf{d}_i \leftarrow \mathcal{P}(\mathbf{d}_{i-1} - \eta \mathbf{u})$ 

16:   return  $\mathbf{d}_k$ 

```

Algorithm 8 Projection to mean. This function seeks to find the vector \mathbf{v}' s.t. the mean of \mathbf{v}' is a target value and the distance of \mathbf{v}' to the original vector \mathbf{v} is minimized. Without bounds, this could simply be accomplished by distributing the difference between the current and target means evenly among each dimension (line 10). However, this adjustment may move some elements of \mathbf{v} outside of the bounds $[0,1]$; imposing these bounds in turn changes the mean. An iterative optimization is this required. The algorithm terminates when the mean is within a tolerance or it has performed too many iterations.

```

1:  $n := \text{dimensions}$  (7)

2:  $\mathbf{v} := \text{point to project, } \in \mathbb{R}^n$ 

3:  $\mu^* := \text{target mean}$  (0.4)

4:  $\epsilon := \text{error tolerance}$  ( $1 \times 10^{-6}$ )

5:  $i_{max} := \text{max iterations}$  (100)

6: procedure  $\mathcal{P}_m(\mathbf{v})$ 

7:    $\Delta_\mu \leftarrow \text{mean}(\mathbf{v}) - \mu^*$ 

8:   while  $|\Delta_\mu| > \epsilon$  and  $i < i_{max}$  do

9:     for  $j = 1..n$  do

10:        $\mathbf{v}_j \leftarrow \mathbf{v}_j - \frac{1}{n}\Delta_\mu$ 

11:        $\mathbf{v}_j \leftarrow \text{max}(0, \text{min}(1, \mathbf{v}_j))$ 

12:        $\Delta_\mu \leftarrow \text{mean}(\mathbf{v}) - \mu^*$ 

13:        $i \leftarrow i + 1$ 

14:   return  $\mathbf{v}$ 

```

Algorithm 9 Projection to range. This function simply performs linear interpolation to map a vector from its original range to the interval $[0,1]$.

```

1:  $n := \text{dimensions}$  (7)

2:  $\mathbf{v} := \text{point to project, } \in \mathbb{R}^n$ 

3: procedure  $\mathcal{P}_r(\mathbf{v})$ 

4:    $v_{min} \leftarrow \mathbf{min}(\mathbf{v})$ 

5:    $v_{max} \leftarrow \mathbf{max}(\mathbf{v})$ 

6:   for  $i = 1..n$  do

7:      $\mathbf{v}_i \leftarrow (\mathbf{v}_i - v_{min}) / (v_{max} - v_{min})$ 

8:   return  $\mathbf{v}$ 

```

12.2 Experimental Design

Each of the participants (the same as those for Exp. 5) completed 20 trials for Exp. 2, which were seamlessly interleaved with the Exp. 5 trials (see Section 10.2). However, different from Exp. 5, the two charts in each trial had the identical summary statistic—there was no *correct* answer (which amounts to the titer value being 0 for all trials). To participants, these trials would seem just like very difficult trials in the same experiment. Like Exp. 5, task-integral factors (*ink area* for *MaxMean*; *min bar* and *max bar* for *MaxRange*) were controlled and balanced between sides (see Section 11.2).

Eight participants for each task (*MaxMean*, *MaxRange*) started from random initializations. Each of the subsequent (35, 34) participants built on a previous result, adding an epoch of optimization, and creating threads of up to 5 epochs. From

these (43, 42) results, we chose (20, 20) for evaluation with subsequent participants, using the participants who performed best at Exp. 5 (lowest final control titer) as a filtering criteria. The remaining (21, 23) participants were shown the final charts of each of these (20, 20) participants compared to random charts. Thus each of these (21, 23) participants saw each of the (20, 20) charts once and only once, and each of the (20, 20) charts was evaluated (21, 23) times.

12.3 Analysis

We focus our analysis on 4 charts for each task that were optimized across 5 epochs and whose final charts were evaluated by other participants. The charts, denoted by \mathcal{M}_i and \mathcal{R}_j ($i, j \in \{1, \dots, 4\}$) for the two tasks, respectively, can be seen next to their random initializations in Figs. 12.1 and 12.2. We performed both quantitative analysis and qualitative visual inspection for these results. To see if the optimized charts reflected the properties of the tested proxies in Exp. 5, we computed the tested proxies from the charts and compared them to a random guessing simulation. The simulation used the same algorithm as initialization, performed 1,000 times with 100 guessing trials (simulating 20 trials per participant \times 5 participants). We then computed median and MAD from the simulation for comparison. We also computed the ratio that a final optimized chart was selected by a participant for that task in the validation trials.

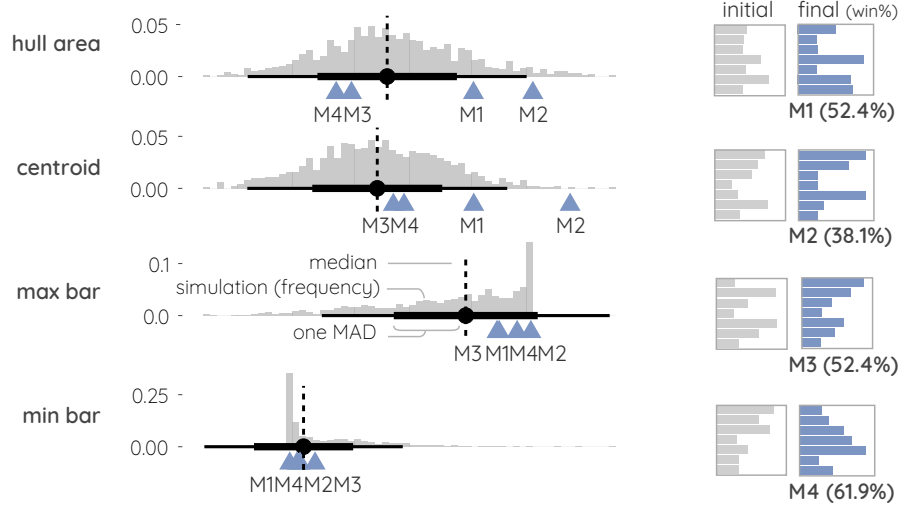


Figure 12.1: **The results of Experiment 6 (\mathcal{H}_1 and \mathcal{H}_2) for MaxMean .** Left, histograms (gray) show distributions of proxy values for charts produced by random choices rather than human experiments, with proxy values for the human-optimized charts (M1–M4) plotted below for comparison. Right, the initializations against the final charts, with percentages indicating how often these charts were actually chosen over random charts in subsequent validation trials.

12.4 Results

Our observations from Experiment 6 are as follows.

MaxMean (Fig. 12.1) We found that \mathcal{M}_1 and \mathcal{M}_2 are at least one MAD away from the median of the random guessing results for *centroid* and *hull area*, and half for *max bar*. In the validation trials, none of the final charts were selected by participants higher than chance (50%). In particular, in \mathcal{M}_1 and \mathcal{M}_2 , the bars have been pushed toward the extrema. We can conjecture that the prominence of the larger bars causes them to carry more weight, increasing the perceived mean. In \mathcal{M}_3 and \mathcal{M}_4 , there are staircase patterns, which may suggest a proxy related to slope.

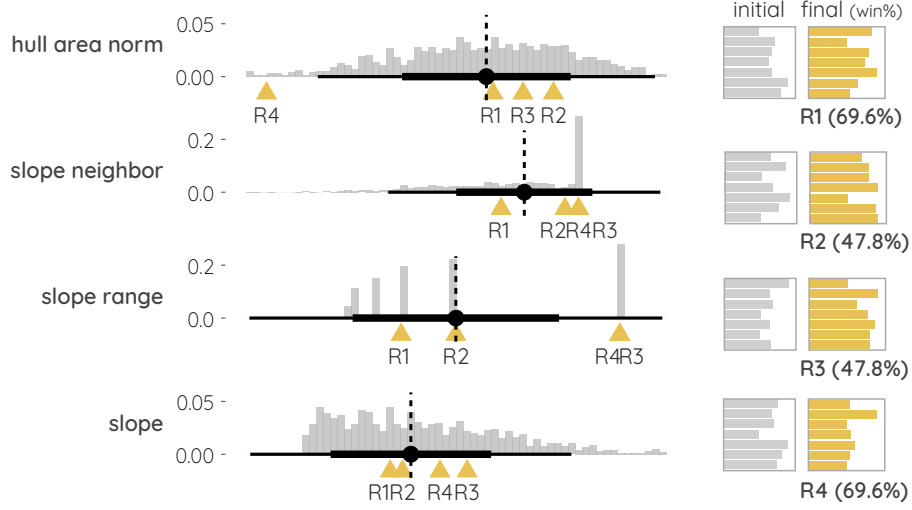


Figure 12.2: **The results of Experiment 6** (\mathcal{H}_1 and \mathcal{H}_2) for *MaxRange*. Left, histograms (gray) show distributions of proxy values for charts produced by random choices rather than human experiments, with proxy values for the human-optimized charts (R1–R4) plotted below for comparison. Right, the initializations against the final charts, with percentages indicating how often these charts were actually chosen over random charts in subsequent validation trials.

MaxRange (Fig. 12.2) We found that \mathcal{R}_3 and \mathcal{R}_4 seem to suggest *slope range*. However, \mathcal{R}_4 is about two MADs from the median of *hull area norm* in the negative direction, slightly suggesting against this *hull area norm*; and \mathcal{R}_1 seems to suggest against most of the proposed proxies. In the validation trials, \mathcal{R}_1 and \mathcal{R}_4 are well above chance (50%), the very similar charts \mathcal{R}_2 and \mathcal{R}_3 are slightly below. This discrepancy could be an effect of individual differences. In \mathcal{R}_3 and \mathcal{R}_4 , there is an inverse of this motif: the maximum is flanked by either the minimum or bars close to it. We expect both of these motifs should correspond with *slope range* and *slope neighbor*. Turning to range, \mathcal{R}_1 and \mathcal{R}_2 appear to have “notches,” in which the shortest bar is flanked by bars near the maximum. We conjecture that this juxtaposition simplifies extraction of

the range. This motif may not make the range appear larger, but easier to estimate, making it more attractive in a forced-choice task.

12.5 Discussion

For *MaxMean*, there is an elegant symmetry to the results of the theory-driven and approaches, in that the charts generated *de novo* (Experiment 6) had, in some cases, similar appearance to the charts specifically optimized for centroids (Experiment 5), and in all cases had relatively high centroid proxy values. While this is far from conclusive evidence, the centroid proxy appears to be the most plausible of the ones we tested for estimation of *MaxMean*. However, this is after controlling for the reductive *ink area* proxy, which is likely the primary mode of estimation when available.

For *MaxRange*, the picture is less clear for both experiments. However, the motifs seen in data-driven *MaxRange* charts may offer lessons for designer further proxies; namely that they account for large adjacent differences in bars (which would create the notches or spikes) that are not necessarily the global minima or maxima of the charts.

We believe this work is only scratching the surface of what is possible with the data-driven approach. We ran relatively low numbers of iterations and initializations, and thus could potentially see patterns better simply by obtaining more data. We also perform only rudimentary analyses, leaving probabilistic evidence of the potency of optimized charts for future study.

Chapter 13: Discussion

Our work lies at the intersection of Data Visualization and Perceptual Psychology, and thus will be of interest to both fields. Its implications for Data Visualization can be further broken down into those relevant to practitioners (i.e. designers of visualizations) and those relevant to researchers.

13.1 Implications for Data Visualization Practice

Ultimately, the goal of Data Visualization research is to improve visualizations in the real world. Many of our earlier experiments on arrangements offer guidance for designers of visualizations that could help to do just that. While our subsequent study of perceptual proxies is perhaps not mature enough to directly impact visualization design, it does at least suggest that the idea of adversarial visualizations is something designers should be aware of.

13.1.1 Design Guidance

Much of this work revolved around filling out the “cube” in Figure 1.1 with empirical study. Though tied closely to specific encodings and tasks, the results of these experiments do provide straightforward design guidance, if limited in scope:

- **Superposition of datasets aids detection of differences.** This has been suggested previously (especially by Gleicher et al. [21]), but is now supported by experimental evidence. When a viewer needs to extract more holistic values (such as means or ranges) from each dataset, though, superposition can be a hindrance.
- **Symmetry has value.** At least for bars, symmetry can help both to emphasize subtle differences in highly correlated data (as seen in population pyramids) and to detect the level of correlation between two datasets. An obvious drawback of symmetry, however, is the limitation to two datasets.
- **Animation can do more than direct attention.** Gleicher et al. [21] suggested that animation could represent a type of explicit encoding of difference, in terms of velocity. We show that this encoding can be effective, and even more so than static visualizations in some cases. Of course, the ephemeral nature of animation may make this difficult to take advantage of in practice.
- **The best layout of small multiples depends on the task.** When making comparisons using small multiples of bar charts, designers have a choice between aligning the baselines (stacking vertically) or aligning bar heights (side-by-side). We show that neither is superior, but that they support different tasks; stacking is better for comparing means and ranges, while side-by-side is better for determining individual bar differences or overall similarity. Note that in all cases these results assume horizontal bars, as we did not test vertical bars.

13.1.2 Adversarial Visualizations and Deception

Our findings may suggest that when a visualization is precisely designed and applied for a specific task, it is possible that participants will be misled simply by virtue of the data. In a way, this is a corollary to Anscombe’s quartet where even a correct (even the “right”) visualization for a specific dataset can be misleading. This hints at some of the “black hat” visualization work discussed by Correll and Heer [78], where it is useful to start to think about visualization in the language of computer security, and where a particular visualization can be open to unintentional (or malicious) attacks even with the best of intentions. However, our efforts to skew perception along these vectors for the sake of investigation have shown that, in practice, this is quite difficult, and likely to be subtle if successful. A malicious designer would thus have many paths of lower resistance [34].

Still, being aware of this problem is the first step towards addressing it. In the short term, establishing the preferred perceptual proxies for not just individuals, but also populations, may allow us to pinpoint situations where unfortunate (or intentional) configurations of data may lead to incorrect perceptions. In the longer term, the perceptual proxies we have investigated here may become the building blocks for perceptual frameworks that are capable of assessing any given visual representation and dataset, and report on the data loss inherent for different subsets of the population.

13.2 Implications for Data Visualization Research

In addition to the results of our experiments, we have also discussed experimental frameworks that push the boundaries of how Data Visualization research is performed. These methods will likely be of interest to other researchers.

- **The Staircase Method translates to Data Visualization.** This framework is often used in psychophysics, which studies elementary perceptual processes. We found it also worked well for the somewhat higher-level processes of performing basic tasks with simple data visualizations. This could be important for Data Visualization research in the future, since crowdsourcing is becoming an ever-more common choice for experiments—by adjusting difficulty dynamically, the Staircase Method helps avoid noise from the variations in experimental setups (e.g. display size and brightness) that are typical of crowdsourcing.

- **Adversarial visualizations can disentangle correlated phenomena.**

The difficulty in studying Perceptual Proxies is that, with typical data, people would make mostly the same choices whether they were using a proxy or computing the real value. We showed that optimization of datasets to be “adversarial” (in our case using simulated annealing) can help to ascribe responses to one process or another. We expect this methodology to be useful for further study of perceptual proxies, and potentially for other types of Data Visualization experiments.

- **Black-box optimization could help characterize perceptual functions.**

While black-box optimization has been used with human judgments before, for example to learn user preferences, here we use it to learn something about how humans are making those judgments. By optimizing the *deceptiveness* of a chart, we seek to learn how humans are deceived, and thus what processes we may be using. Though our results from black-box optimizations are very preliminary and largely qualitative, we think this method has much potential for learning Perceptual Proxies and potentially many other neurological processes.

13.3 Implications for Perceptual Psychology

How the brain translates charts from images on the retina into more abstract conceptual relationships is still largely a mystery. Frameworks such as Pinker’s Theory of Graph Comprehension [121], however, do provide plausible mechanisms that make testable predictions. Though in this work we use predictions of Perceptual Psychology largely as a means to the end of creating more effective charts, our work can also be seen as testing some of those predictions, providing valuable information to Perceptual Psychologists in return. Here we will demonstrate this value using Pinker’s proposed model as a framework. Note, however, that other interpretations based on different theories of graph comprehension are possible.

Briefly, in Pinker’s model, a visualization is represented in the brain by a hierarchical “scene graph,” with nodes corresponding to perceptual elements ranging

from low levels (e.g. lengths and shapes) to higher levels (e.g. Gestalt organizations and relationships). Observed attributes of these nodes are formalized as “predicates,” which attach specific values to them (such as the length of an individual bar or the fact that a group of bars makes a descending staircase). Perceptual Proxies can be thought of as predicates of mid- or high-level nodes in the scene graph. Our results have several implications for this analogy.

- **Extraction of statistics can be non-compositional.** If the brain always performed computations on chart data by extracting individual values (e.g. the length of each bar in a bar chart) and then performing calculations with them, we would not expect the relative positioning of charts (i.e. *arrangements*) to affect computations. To the contrary, our results from Experiments 1–4 show that comparative arrangement can make significant differences in how quickly and accurately comparisons are performed. This phenomenon in its own right is compelling evidence that, at least in some settings, Perceptual Proxies are used, rather than compositions of lower-level perceptual operations.
- **Perceptual Proxies can correspond to Gestalt phenomena.** In Experiment 2 we showed that the similarity of data in two bar charts can be more easily perceived when that similarity corresponds to the level of bilateral symmetry in the overall scene.¹ It would be hard to explain this result if not for the visual system’s natural inclination to recognize this type of symmetry

¹Since the applications for symmetry in Data Visualization practice are rather arcane, this is an example of a result that may actually be of more value in the context of Perceptual Psychology.

(see §2.2.2). The representation of visualizations as Gestalt elements could also explain why superposition (or “overlaid” displays) hinders extraction of summary statistics such as the mean or range (as seen in Experiments 3 and 4, respectively), since interspersing bars from multiple charts would interfere with the visual system’s ability to recognize each dataset as a contiguous object. Finally, though not a conclusive result, we show some evidence that the centroid of bars in a bar chart can be used as a proxy to estimate their mean, which further hints at the representation of charts at various levels in a hierarchical scene graph, with Gestalt properties attached to nodes as predicates.

- **Perceptual Proxies can be attached to abstract meanings.** The attachment of visual properties to nodes in a scene graph is only part of the picture of graph comprehension—the brain still needs a bridge from values to the more abstract concepts that a graph is representing, for example the fact that the length of a bar corresponds to, say, the number barrels of oil produced in a given month. In Pinker’s model, these relationships are called “message flags,” and together with the scene graph, they form a more complete “schema,” or mental model, that a viewer can apply when a graph is encountered. Pinker posits that these flags can also exist at higher-level nodes, for example the prior knowledge that a bar chart with a wedge-like outline means there is an increasing or decreasing trend. In fact, Pinker further posits that these higher-level flags are what give visualizations much of their power for rapid insights. To continue with the example of bilateral symmetry, we did not

tell participants to pick the more symmetrical image, but rather to interpret images as bar charts and choose the pair with more similar *data*. In order to bring their faculties for detecting symmetry to bear within this task (which is what the evidence suggests they were able to do), participants must have, at some level, made the inference that this basic Gestalt perception had the more abstract meaning of similarity in the context of the charts. This is in line with what the existence of message flags would predict. Likewise, animation was likely helpful for conveying subtle differences in Experiment 1 because of the encoding of those differences as velocity, which the visual system is good at estimating (see §2.2.3). Interestingly, in both these cases, participants were not likely to have used these particular “message flags” before, supporting the idea that graph comprehension schema can be acquired and modified with experience.

13.4 Limitations

While our approaches revealed many interesting findings, they are limited in many ways.

- **Experimental context:** While highly controlled experimental conditions are crucial to empirical evaluation, they can often be at odds with the ecological validity of the results [122]. In this case, for example, our studies show that both mirror symmetry and animation can be beneficial in certain, specific contexts, but do those benefits extend to applications in the real world?

- **Choice of proxies:** For our theory-driven approach, though we added proxies to those used in previous studies, we still cannot claim to have anything approaching an exhaustive list, nor can we claim strong motivations for testing these particular proxies. Additionally, we intentionally omit some proxies that are either directly connected to their corresponding summary statistics or highly correlated with chosen proxies, which would be difficult to control independently, and thus to measure. While this necessarily limits the conclusions we can draw about specific proxies that lie within broader classes, we believe probing these few representatives is a necessary first step towards disentangling the myriad of proxies that have been proposed, and are yet to be conceived of.
- **Visual Modeling:** The proxies implemented here did not use a computer visual system to “look at” pixels of a chart’s visual features and parse those pixels into values. We used the actual data values to generate models of these perceptual proxies. The value of this approach is that if we can determine the properties of the data and arrangements that lend themselves to particular proxies for comparison, then a potential application of this approach is that an automated visualization system would only need know the data values and the designer’s desired comparison to construct the mark and arrangement to support that comparison. In other words, these proxies do not directly take into account limitations in perceptual visual acuity, or the capacity limitations of attention and memory.

Chapter 14: Future Work

Though this work answers some questions, it poses many as well. However, both the results we present and the experimental methods we have developed provide many potential avenues for extending it in the future.

14.1 Continuing to Solve the Cube

Though Experiment 1 included several encodings, our work largely focuses on bar charts, due to their ubiquity in real visualizations, and combination of position and length encodings. Bar charts are clearly flexible visual representations in that they support both global and focal visual comparison, and it is clear from the richness of our results that this limitation did not restrict the complexity of the performance results. As we have discussed, it is likely infeasible to test every possible combination of encoding, task, arrangement, etc. to provide design guidance. Nevertheless, it will clearly be important to continue to fill out the “cube” proposed in Ch. 1, both to test the robustness of the cube model and to provide more data for the enterprise of searching for candidate perceptual proxies for visualization tasks. For example, the bar charts in the present work and that of Ondov et al. [32] were horizontally extended, an increasingly common design [123]. Other variants even of

bar charts might reveal the use of different proxies for comparison.

14.2 Generating New Candidate Proxies

Future work should generate more, and more sophisticated, proxies (including combinations of proxies, and eventually, predictions for who will use which, and when). We generated proxies with a combination of intuition and consultation with the perceptual psychology literature, including a strong influence of the literature on focal vs. global processing modes in vision. Our list is by no means exhaustive, and identifying new candidates will be a creative process that, like hypothesis generation across the rest of science, relies on engaging a diverse group of people with different types of background knowledge across both the perception and data visualization communities. A brute force approach would be to generate the full space of mathematically possible pairwise and set-wise proxies. Another route could be based on interviews with viewers engaged in a particular task, to see which aspects of their proxies might be consciously verbalized.

14.3 Proxies Cubed

Just as we proposed perceptual proxies as a reasoning framework to raise the abstraction level and explain all of these phenomena in one fell swoop, must we also endeavor to understand the relationship between different proxies for different visualizations, layouts, and tasks. Put differently, it is highly unlikely that the visual system has developed specialized “programs” (or proxies) for every conceivable vi-

sual representation. It is more likely that there are clear commonalities between the proxies used for different tasks, and moreover that specific individuals have specific affinities for various such proxies. In fact, our population analysis provides some support for this hypothesis. This would mean that a fruitful gradient to optimize for future work would be to try to identify and generalize perceptual proxies across different visualizations and tasks, essentially exploring the “cube” of proxy space.

14.4 How Might Viewers Choose Proxies?

Proxies could be learned, or at least encouraged, from prior experience. For example, scatterplots are often used to communicate a single statistic (correlation) of a set for which precision is important. A viewer seeing a scatterplot will likely develop the analytic goal of perceiving correlation, which should be more likely to trigger analysis of the proxies available in the scatterplot visualization to calculate correlation [121]. These kinds of contextual assumptions may be at play for bar charts and others as well, and investigating this would be important for fully understanding how proxies are used.

14.5 Automated Systems

Of course, one of main goals of exploring proxies is to improve actual visualizations. While design guidance is part of this, a potentially more impactful route may be to take advantage of recent developments in automated visualization recommender systems [124–127]. Much of this work stems from the ideas of Mackin-

lay’s formal *composition algebra* [128], which allows charts to be optimized using definitions of *expressiveness* and *effectiveness*, the latter being largely based on the perceptual studies of Cleveland & McGill [2]. This type of system could be naturally extended to include rules for comparative displays that incorporate our empirical evidence for Experiments 1-4 (and future cells of the “cube”) in their definitions of *effectiveness*. Additionally, as we gain a better understanding of proxies, encodings of effectiveness could become simple models of the vision system, effectively “seeing” the data. This would allow them to optimize encodings and arrangements based on the task required, or to warn designers when a chart might be deceptive.

Chapter 15: Conclusion

We began this work with a case study illustrating the importance of comparison in data visualization. Though it involved a very specific type of visualization and a single domain, it emphasized that not all modes of visual comparison are equal, and that the best mode may depend on the task at hand. This led us to more thoroughly investigate the factors at play in creating effective comparative visualizations. Experiments 1–4 provided empirical evidence for which arrangements best support certain tasks. However, they also showed that making recommendations will not be as straightforward as for elementary encodings, as the visual operations for comparison do not seem to be strictly compositional. This led to the question of how the visual system actually does perform these comparative operations. Experiments 5 and 6 approached this question from two different directions; that is, we used both theory-driven (Experiment 5) and data-driven (Experiment 6) approaches to seek evidence that participants might have used perceptual proxies in mean and range tasks. Experiment 5 explored “proxy space” by carefully optimizing datasets based on predefined proxies, whereas Experiment 6 explored “data space,” doing away with any preconceived notions of chart characteristics to optimize charts directly by their deceptiveness. The meeting point of these two experiments is the

evidence that participants might have used certain and the same proxies in both experiments. For example, for the *MaxMean* task, both experiments suggest that participants might have used *centroid* as a proxy.

As a whole, this work has many implications, but they can be broadly divided into those that are theoretical and those that are practical. Our initial experiments comparing visual arrangements offer some directly applicable guidance for designers wishing to maximize the efficacy of such displays. However, they also provide some evidence for the theory that cognition of higher-level properties of charts is not simply compositional of lower-level perceptual operations. Our adversarial experiments with perceptual proxies take a step toward understanding what those cognitive operations actually are. This, in turn, is a step toward providing guidance for the design of comparative displays that is general enough to avoid empirical evaluation of every combination of encoding, task, arrangement, and other potential factors. Finally, another aspect of our work that is useful to the visualization community is the methodological framework we have devised to test these phenomena. We hope to see future studies in visualization use similar reactive testing frameworks such as ours to empirically derive increasingly more complex visual phenomena.

Bibliography

- [1] Jacques Bertin. *Semiology of Graphics*. University of Wisconsin Press, Madison, Wisconsin, 1983.
- [2] William S. Cleveland and Robert McGill. Graphical perception: Theory, experimentation and application to the development of graphical methods. *Journal of the American Statistical Association*, 79(387):531–554, September 1984.
- [3] Jeffrey Heer and Michael Bostock. Declarative language design for interactive visualization. *IEEE Transactions on Visualization and Computer Graphics*, 16(6):1149–1156, 2010.
- [4] Matthew Brehmer, Jocelyn Ng, Kevin Tate, and Tamara Munzner. Matches, mismatches, and methods: Multiple-view workflows for energy portfolio analysis. *IEEE Transactions on Visualization and Computer Graphics*, 22(1):449–458, 2016.
- [5] Michael Gleicher. Considerations for visualizing comparison. *IEEE Transactions on Visualization and Computer Graphics*, 24(1):413–423, 2017.
- [6] George Robertson, Roland Fernandez, Danyel Fisher, Bongshin Lee, and John Stasko. Effectiveness of animation in trend visualization. *IEEE Transactions on Visualization and Computer Graphics*, 14(6):1325–1332, 2008.
- [7] Robert A. Amar and John T. Stasko. Knowledge precepts for design and evaluation of information visualizations. *IEEE Transactions on Visualization and Computer Graphics*, 11(4):432–442, July 2005.
- [8] Steven L Franconeri. The nature and status of visual resources. *Oxford Handbook of Cognitive Psychology*, 8481:147–162, 2013.
- [9] D. Simkin and R. Hastie. An information-processing analysis of graph perception. *Journal of the American Statistical Association*, 82(398):454–465, 1987.

- [10] Younghoon Kim and Jeffrey Heer. Assessing effects of task and data distribution on the effectiveness of visual encodings. In *Computer Graphics Forum*, volume 37, pages 157–167. Wiley Online Library, 2018.
- [11] Robert Amar and John Stasko. Best paper: A knowledge task-based framework for design and evaluation of information visualizations. In *IEEE Symposium on Information Visualization*, pages 143–150. IEEE, 2004.
- [12] Brian D. Ondov, Nicholas H. Bergman, and Adam M. Phillippy. Interactive metagenomic visualization in a web browser. *BMC Bioinformatics*, 12(385), Sep 2011.
- [13] J. Stasko and E. Zhang. Focus+context display and navigation techniques for enhancing radial, space-filling hierarchy visualizations. *Proceedings of the IEEE Symposium on Information Visualization*, pages 57–65, 2000.
- [14] Edward Tufte. *Envisioning Information*. Graphics Press, Cheshire, CT, USA, 1990.
- [15] Adam Barnas and Adam Greenberg. Visual field meridians modulate the re-allocation of object-based attention. *Attention, Perception, & Psychophysics*, 78(7):1985–1997, 05 2016.
- [16] Bryan Matlen, Dedre Gentner, and Steve Franconeri. Structure mapping in visual comparison: Embodied correspondence lines? In *Proceedings of the Annual Meeting of the Cognitive Science Society*, 2014.
- [17] Waqas Javed and Niklas Elmqvist. Exploring the design space of composite visualization. In *Proceedings of the IEEE Pacific Symposium on Visualization*, pages 1–8, 2012.
- [18] Ken Nakayama. Biological image motion processing: a review. *Vision research*, 25(5):625–660, 1985.
- [19] Brian R Levinthal and Steven L Franconeri. Common-fate grouping as feature selection. *Psychological science*, 22(9):1132–1137, 2011.
- [20] Johan Wagemans. Characteristics and models of human symmetry detection. *Trends in Cognitive Sciences*, 1(9):346–352, 1997.
- [21] Michael Gleicher, Danielle Albers, Rick Walker, Ilir Jusufi, Charles D. Hansen, and Jonathan C. Roberts. Visual comparison for information visualization. *Information Visualization*, 10(4):289–309, October 2011.
- [22] S. Limoges, C. Ware, and W. Knight. Displaying correlations using position, motion, point size or point colour. In *Proceedings of Graphics Interface*, pages 262–265, Toronto, Ontario, Canada, 1989. Canadian Man-Computer Communications Society.

- [23] Fanny Chevalier, Pierre Dragicevic, and Steven Franconeri. The not-so-staggering effect of staggered animated transitions on visual tracking. *IEEE transactions on visualization and computer graphics*, 20(12):2241–2250, 2014.
- [24] F. Yang, L. Harrison, R. A. Rensink, S. Franconeri, and R. Chang. Correlation judgment and visualization features: A comparative study. *IEEE Transactions on Visualization and Computer Graphics*, PP(99):1–1, 2018.
- [25] L. Yuan, S. Haroz, and S. Franconeri. Perceptual proxies for extracting averages in data visualizations. *Psychonomic Bulletin & Review*, 26:669–676, 2019.
- [26] Nicole Jardine, Brian D Ondov, Niklas Elmqvist, and Steven Franconeri. The perceptual proxies of visual comparison. *IEEE transactions on visualization and computer graphics*, 26(1):1012–1021, 2019.
- [27] Nilesh Dalvi, Pedro Domingos, Sumit Sanghai, and Deepak Verma. Adversarial classification. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 99–108, 2004.
- [28] Ling Huang, Anthony D Joseph, Blaine Nelson, Benjamin IP Rubinstein, and J Doug Tygar. Adversarial machine learning. In *Proceedings of the 4th ACM workshop on Security and artificial intelligence*, pages 43–58, 2011.
- [29] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.
- [30] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial machine learning at scale. *arXiv preprint arXiv:1611.01236*, 2016.
- [31] Richard Langton Gregory. Perceptual illusions and brain models. *Proceedings of the Royal Society of London. Series B. Biological Sciences*, 171(1024):279–296, 1968.
- [32] Brian D. Ondov, Nicole Jardine, Niklas Elmqvist, and Steven Franconeri. Face to face: Evaluating visual comparison. *IEEE Transactions on Visualization and Computer Graphics*, 25(1):861–871, 2019.
- [33] Brian D Ondov, Fumeng Yang, Matthew Kay, Niklas Elmqvist, and Steven Franconeri. Revealing perceptual proxies with adversarial examples. *IEEE Transactions on Visualization and Computer Graphics*, 2020.
- [34] Edward R. Tufte. *The Visual Display of Quantitative Information*. Graphics Press, Cheshire, CT, USA, 1986.
- [35] John Duncan and Glyn W. Humphreys. Visual search and stimulus similarity. *Psychological Review*, 96(3):433–458, 08 1989.

- [36] Laurent Itti and Christof Koch. A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research*, 40(10):1489–1506, 2000.
- [37] Anne M. Treisman and Garry Gelade. A feature-integration theory of attention. *Cognitive Psychology*, 12(1):97–136, 1980.
- [38] Jeremy Wolfe and Todd Horowitz. Five factors that guide attention in visual search. *Nature Human Behaviour*, 1:0058, 03 2017.
- [39] David Borland and Russell M. Taylor II. Rainbow color map (still) considered harmful. *IEEE Computer Graphics & Applications*, 27(2):14–17, March 2007.
- [40] Samuel Silva, Beatriz Sousa Santos, and Joaquim Madeira. Using color in visualization: A survey. *Computers & Graphics*, 35(2):320–333, 2011.
- [41] Liang Zhou and Charles D Hansen. A survey of colormaps in visualization. *IEEE Transactions on Visualization and Computer Graphics*, 22(8):2051–2069, 08 2016.
- [42] Mehmet Adil Yalçın, Niklas Elmqvist, and Benjamin B. Bederson. Raising the bars: Evaluating treemaps vs. wrapped bars for dense visualization of sorted numeric data. In *Proceedings of the Graphics Interface Conference*, pages 41–49. Canadian Human-Computer Communications Society / ACM, 2017.
- [43] Danielle Albers Szafr. Modeling color difference for visualization design. *IEEE Transactions on Visualization and Computer Graphics*, 24:392–401, 2017.
- [44] Jacques Bertin. *Graphics and Graphic Information Processing*. Walter de Gruyter, Berlin, 1981.
- [45] Robert Amar, James Eagan, and John Stasko. Low-level components of analytic activity in information visualization. In *Proceedings of the IEEE Symposium on Information Visualization*, pages 111–117, Washington, DC, USA, 2005. IEEE Computer Society.
- [46] Zening Qu and Jessica Hullman. Keeping multiple views consistent: Constraints, validations, and exceptions in visualization authoring. *IEEE Transactions on Visualization and Computer Graphics*, 24:468–477.
- [47] Zening Qu and Jessica Hullman. Evaluating visualization sets: Trade-offs between local effectiveness and global consistency. In *Proceedings of the Sixth Workshop on Beyond Time and Errors on Novel Evaluation Methods for Visualization*, BELIV ’16, pages 44–52, New York, NY, USA, 2016. ACM.
- [48] Gapminder. <http://www.gapminder.org>.

- [49] Jeffrey Heer, Nicholas Kong, and Maneesh Agrawala. Sizing the horizon: The effects of chart size and layering on the graphical perception of time series visualizations. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '09, pages 1303–1312, New York, NY, USA, 2009. ACM.
- [50] Waqas Javed, Bryan McDonnell, and Niklas Elmqvist. Graphical perception of multiple time series. *IEEE Transactions on Visualization and Computer Graphics*, 16(6):927–934, 2010.
- [51] Timothy F Brady, Talia Konkle, and George A Alvarez. A review of visual memory capacity: Beyond individual items and toward structured representations. *Journal of vision*, 11(5):4–4, 2011.
- [52] Daniel Simons. Current approaches to change blindness. *Visual Cognition*, 7(1–3):1–15, 01 2000.
- [53] Bela Julesz. *Foundations of Cyclopean Perception*. Chicago University Press, Chicago, Illinois, 1971.
- [54] H. B. Barlow and B. C. Reeves. The versatility and absolute efficiency of detecting mirror symmetry in random dot displays. *Vision Research*, 19(7):783–793, 1979.
- [55] Matthias Treder. Behind the looking-glass: A review on human symmetry perception. *Symmetry*, 2(3):1510–1543, 09 2010.
- [56] Michael C. Corballis and Carlos E. Roldan. On the perception of symmetrical and repeated patterns. *Perception & Psychophysics*, 16(1):136–142, Jan 1974.
- [57] Simona Korenjak-Ćerne, Nataša Kejžar, and Vladimir Batagelj. Clustering of population pyramids. *Informatica*, 32(2), 2008.
- [58] Richard Langton Gregory. *Eye and Brain*. McGraw-Hill, New York, New York, 1973.
- [59] Jason Haberman and David Whitney. Ensemble perception: Summarizing the scene and broadening the limits of visual processing. *From perception to consciousness: Searching with Anne Treisman*, pages 339–349, 2012.
- [60] Danielle Albers Szafr, Steve Haroz, Michael Gleicher, and Steven Franconeri. Four types of ensemble coding in data visualizations. *Journal of Vision*, 16(5):11–11, 2016.
- [61] Jason M Scimeca and Steven L Franconeri. Selecting and tracking multiple objects. *Wiley Interdisciplinary Reviews: Cognitive Science*, 6(2):109–118, 2015.

- [62] Yangqing Xu and Steven L Franconeri. Capacity for visual features in mental rotation. *Psychological science*, 26(8):1241–1251, 2015.
- [63] Barbara Tversky, Julie Bauer Morrison, and Mireille Betrancourt. Animation: can it facilitate? *International Journal of Human-Computer Studies*, 57(4):247–262, 2002.
- [64] Fanny Chevalier, Nathalie Henry Riche, Catherine Plaisant, Amira Chalbi, and Christophe Hurter. Animations 25 years later: New roles and opportunities. In *Proceedings of the ACM Conference on Advanced Visual Interfaces*, pages 280–287. ACM, 2016.
- [65] Daniel Archambault, Helen Purchase, and Bruno Pinaud. Animation, small multiples, and the effect of mental map preservation in dynamic graphs. *IEEE Transactions on Visualization and Computer Graphics*, 17(4):539–552.
- [66] Fan Du, Nan Cao, Jian Zhao, and Yu-Ru Lin. Trajectory bundling for animated transitions. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*, pages 259–268, 2015.
- [67] Amy L. Griffin, Alan M. MacEachren, Frank Hardisty, Erik Steiner, and Bonan Li. A comparison of animated maps with static small-multiple maps for visually identifying space-time clusters. *Annals of the Association of American Geographers*, 96(4):740–753, 2006.
- [68] Jeffrey Heer and George Robertson. Animated transitions in statistical data graphics. *IEEE Transactions on Visualization and Computer Graphics*, 13(6):1240–1247, November 2007.
- [69] David Marr. *Vision: A Computational Investigation Into the Human Representation and Processing of Visual Information*. W. H. Freeman and Company, San Francisco, CA, USA, 1982.
- [70] J. H. Larkin and H. A. Simon. Why a diagram is (sometimes) worth ten thousand words. *Cognitive Science*, 11(1):65–100, 1987.
- [71] C. M. Carswell. Choosing specifiers: an evaluation of the basic tasks model of graphical perception. *Human Factors*, 34(5):535–554, 1992.
- [72] R. M. Ratwani, J. G. Trafton, and D. A. Boehm-Davis. Thinking graphically: Connecting vision and cognition during graph comprehension. *Journal of Experimental Psychology: Applied*, 14(1):36–49, 2008.
- [73] Lane Harrison, Fumeng Yang, Steven Franconeri, and Remco Chang. Ranking visualizations of correlation using Weber’s law. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):1943–1952, 2014.
- [74] Ronald A. Rensink and Gideon Baldridge. The perception of correlation in scatterplots. *Computer Graphics Forum*, 29(3):1203–1210, 2010.

- [75] Jeff Zacks and Barbara Tversky. Bars and lines: A study of graphic communication. *Memory & Cognition*, 27(6):1073–1079, 1999.
- [76] Hadley Wickham, Dianne Cook, Heike Hofmann, and Andreas Buja. Graphical inference for infovis. *IEEE Transactions on Visualization and Computer Graphics*, 16(6):973–979, 2010.
- [77] Anshul Vikram Pandey, Katharina Rall, Margaret L. Satterthwaite, Oded Nov, and Enrico Bertini. How deceptive are deceptive visualizations?: An empirical analysis of common distortion techniques. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*, pages 1469–1478, New York, NY, USA, 2015. ACM.
- [78] Michael Correll and Jeffrey Heer. Black hat visualization. In *Proceedings of the IEEE VIS Workshop on Dealing with Cognitive Biases in Visualisations*, 2017.
- [79] Michael Correll, Mingwei Li, Gordon Kindlmann, and Carlos Scheidegger. Looks good to me: Visualizations as sanity checks. *IEEE Transactions on Visualization and Computer Graphics*, 25(1):830–839, 2018.
- [80] Ben Shneiderman. Tree visualization with tree-maps: 2-D space-filling approach. *ACM Transactions on Graphics*, 11(1):92–99, January 1992.
- [81] Maxime Hebrard and Todd D. Taylor. Metatreemap: An alternative visualization method for displaying metagenomic phylogenetic trees. *PLOS ONE*, 11(6):1–6, 06 2016.
- [82] Willard Cope Brinton. *Graphic Presentation*. Brinton Associates, New York, New York, 1939.
- [83] Florian P Breitwieser and Steven L Salzberg. Pavian: Interactive analysis of metagenomics data for microbiomics and pathogen identification. *bioRxiv*, 06 2016.
- [84] J. Gregory Caporaso, Christian L. Lauber, Elizabeth K. Costello, Donna Berg-Lyons, Antonio Gonzalez, Jesse Stombaugh, Dan Knights, Pawel Gajer, Jacques Ravel, Noah Fierer, Jeffrey I. Gordon, and Rob Knight. Moving pictures of the human microbiome. *Genome Biology*, 12(5):R50, May 2011.
- [85] Miriah Meyer, Bang Wong, Mark Styczynski, Tamara Munzner, and Hanspeter Pfister. Pathline: A tool for comparative functional genomics. In *Computer Graphics Forum*, volume 29, pages 1043–1052. Wiley Online Library, 2010.
- [86] Pierre Dragicevic, Anastasia Bezerianos, Waqas Javed, Niklas Elmqvist, and Jean-Daniel Fekete. Temporal distortion for animated transitions. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*, pages 2009–2018, 2011.

- [87] Michael Bostock, Vadim Ogievetsky, and Jeffrey Heer. D3: data-driven documents. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2301–2309, 2011.
- [88] Dan Murray. *Tableau Your Data!: Fast and Easy Visual Analysis with Tableau Software*. Wiley Publishing, 1st edition, 2013.
- [89] Max Sondag, Bettina Speckmann, and Kevin Verbeek. Stable treemaps via local moves. *IEEE Trans. Vis. Comput. Graph*, 24(1):729–738, 2018.
- [90] Justin Matejka and George Fitzmaurice. Same stats, different graphs: Generating datasets with varied appearance and identical statistics through simulated annealing. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*, pages 1290–1294, New York, NY, USA, 2017. ACM.
- [91] George A Alvarez. Representing multiple objects as an ensemble enhances visual cognition. *Trends in Cognitive Sciences*, 15(3):122–131, 2011.
- [92] David Whitney and Allison Yamanashi Leib. Ensemble perception. *Annual Review of Psychology*, 69:105–129, 2018.
- [93] Heeyoung Choo, Brian R Levinthal, and Steven L Franconeri. Average orientation is more accessible through object boundaries than surface features. *Journal of Experimental Psychology: Human Perception and Performance*, 38(3):585, 2012.
- [94] Zhe Chen. Object-based attention: A tutorial review. *Attention, Perception, & Psychophysics*, 74(5):784–802, 2012.
- [95] D Melcher and E Kowler. Shapes, surfaces and saccades. *Vision Research*, 39(17):2929–2946, 1999.
- [96] Michael King, Glenn E Meyer, John Tangney, and Irving Biederman. Shape constancy and a perceptual bias towards symmetry. *Perception & Psychophysics*, 19(2):129–136, 1976.
- [97] Dennis M. Levi and Stanley A. Klein. Vernier acuity, crowding and amblyopia. *Vision Research*, 25(7):979–991, 1985.
- [98] Laurent Itti, Christof Koch, and Ernst Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (11):1254–1259, 1998.
- [99] Marina G Falletti, Paul Maruff, Alexander Collie, and David G Darby. Practice effects associated with the repeated assessment of cognitive function using the CogState battery at 10-minute, one week and one month test-retest intervals. *Journal of Clinical and Experimental Neuropsychology*, 28(7):1095–1112, 2006.

- [100] Donna M Webster, Linda Richter, and Arie W Kruglanski. On leaping to conclusions when feeling tired: Mental fatigue effects on impressional primacy. *Journal of Experimental Social Psychology*, 32(2):181–195, 1996.
- [101] Scott Kirkpatrick, C Daniel Gelatt, and Mario P Vecchi. Optimization by simulated annealing. *Science*, 220(4598):671–680, 1983.
- [102] George A Gescheider. The classical psychophysical methods. *Psychophysics: the fundamentals*, pages 45–72, 1997.
- [103] Tom N Cornsweet. The staircase-method in psychophysics. *The American Journal of Psychology*, 75(3):485–491, 1962.
- [104] Eric-Jan Wagenmakers, Maarten Marsman, Tahira Jamil, Alexander Ly, Josine Verhagen, Jonathon Love, Ravi Selker, Quentin F Gronau, Martin Šmíra, Sacha Epskamp, et al. Bayesian inference for psychology. Part I: theoretical advantages and practical ramifications. *Psychonomic Bulletin & Review*, 25(1):35–57, 2018.
- [105] Matthew Kay, Gregory L. Nelson, and Eric B. Hekler. Researcher-centered design of statistics: Why Bayesian statistics better fit the culture and incentives of HCI. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*, pages 4521–4532, New York, NY, USA, 2016. ACM.
- [106] Paul-Christian Bürkner et al. brms: An r package for bayesian multilevel models using stan. *Journal of Statistical Software*, 80(1):1–28, 2017.
- [107] Matthew Kay. *ggdist: Visualizations of Distributions and Uncertainty*, 2020. R package version 2.2.0.9000.
- [108] Matthew Kay. *tidybayes: Tidy Data and Geoms for Bayesian Models*, 2020. R package version 2.1.1.9000.
- [109] Stan Development Team. RStan: the R interface to Stan, 2018. R package version 2.18.2.
- [110] Hadley Wickham, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D’Agostino McGowan, Romain François, Garrett Golemund, Alex Hayes, Lionel Henry, Jim Hester, Max Kuhn, Thomas Lin Pedersen, Evan Miller, Stephan Milton Bache, Kirill Müller, Jeroen Ooms, David Robinson, Dana Paige Seidel, Vitalie Spinu, Kohske Takahashi, Davis Vaughan, Claus Wilke, Kara Woo, and Hiroaki Yutani. Welcome to the tidyverse. *Journal of Open Source Software*, 4(43):1686, 2019.
- [111] Alex Kale, Francis Nguyen, Matthew Kay, and Jessica Hullman. Hypothetical outcome plots help untrained observers judge trends in ambiguous data. *IEEE Transactions on Visualization and Computer Graphics*, 25(1):892–902, 2018.

- [112] J Martin Bland and Douglas G Altman. Statistics notes: Measurement error. *British Medical Journal*, 312(7047):1654, 1996.
- [113] Richard McElreath. *Statistical Rethinking: A Bayesian Course with Examples in R and Stan*. CRC Press, 2015.
- [114] Christophe Leys, Christophe Ley, Olivier Klein, Philippe Bernard, and Laurent Licata. Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *Journal of Experimental Social Psychology*, 49(4):764–766, 2013.
- [115] John K Kruschke. Bayesian estimation supersedes the t test. *Journal of Experimental Psychology: General*, 142(2):573, 2013.
- [116] G. N. Wilkinson and C. E. Rogers. Symbolic description of factorial models for analysis of variance. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 22(3):392–399, 1973.
- [117] Michael Fernandes, Logan Walls, Sean Munson, Jessica Hullman, and Matthew Kay. Uncertainty displays using quantile dotplots or cdfs improve transit decision-making. In *Proc. SIGCHI*, pages 144:1–144:12, 2018.
- [118] Jessica Hullman, Matthew Kay, Yea-Seul Kim, and Samana Shrestha. Imagining replications: Graphical prediction & discrete visualizations improve recall & estimation of effect uncertainty. *IEEE TVCG*, 24(1):446–456, 2018.
- [119] Luis Miguel Rios and Nikolaos V Sahinidis. Derivative-free optimization: a review of algorithms and comparison of software implementations. *Journal of Global Optimization*, 56(3):1247–1293, 2013.
- [120] Yisong Yue and Thorsten Joachims. Interactively optimizing information retrieval systems as a dueling bandits problem. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 1201–1208, 2009.
- [121] Steven Pinker. A theory of graph comprehension. In R. Freedle, editor, *Artificial Intelligence and the Future of Testing*, pages 73–126. Lawrence Erlbaum Associates, Inc., Hillsdale, NJ, USA, 1990.
- [122] Niklas Elmqvist and Ji Soo Yi. Patterns for visualization evaluation. *Information Visualization*, 14(3):250–269, 2015.
- [123] Stephen Few. *Now You See It: Simple Visualization Techniques for Quantitative Analysis*. Analytics Press, Oakland, CA, USA, 2009.
- [124] Kanit Wongsuphasawat, Dominik Moritz, Anushka Anand, Jock Mackinlay, Bill Howe, and Jeffrey Heer. Voyager: Exploratory analysis via faceted browsing of visualization recommendations. *IEEE transactions on visualization and computer graphics*, 22(1):649–658, 2015.

- [125] Dominik Moritz, Chenglong Wang, Greg L Nelson, Halden Lin, Adam M Smith, Bill Howe, and Jeffrey Heer. Formalizing visualization design knowledge as constraints: Actionable and extensible models in draco. *IEEE transactions on visualization and computer graphics*, 25(1):438–448, 2018.
- [126] Kanit Wongsuphasawat, Dominik Moritz, Anushka Anand, Jock Mackinlay, Bill Howe, and Jeffrey Heer. Towards a general-purpose query language for visualization recommendation. In *Proceedings of the Workshop on Human-In-the-Loop Data Analytics*, pages 1–6, 2016.
- [127] Kanit Wongsuphasawat, Zening Qu, Dominik Moritz, Riley Chang, Felix Ouk, Anushka Anand, Jock Mackinlay, Bill Howe, and Jeffrey Heer. Voyager 2: Augmenting visual analysis with partial view specifications. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 2648–2659, 2017.
- [128] Jock Mackinlay. Automating the design of graphical presentations of relational information. *Acm Transactions On Graphics (Tog)*, 5(2):110–141, 1986.